

Lecture Notes
in Geoinformation and Cartography

LNG&C

Phaedon Kyriakidis
Diofantos Hadjimitsis
Dimitrios Skarlatos
Ali Mansourian *Editors*

Geospatial Technologies for Local and Regional Development

Proceedings of the 22nd AGILE
Conference on Geographic Information
Science

 Springer

Lecture Notes in Geoinformation and Cartography

Series Editors

William Cartwright, School of Science, RMIT University, Melbourne,
VIC, Australia

Georg Gartner, Department of Geodesy and Geoinformation, Vienna
University of Technology, Wien, Austria

Liqiu Meng, Department of Civil, Geo and Environmental Engineering,
Technische Universität München, München, Germany

Michael P. Peterson, Department of Geography and Geology,
University of Nebraska at Omaha, Omaha, NE, USA

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

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

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

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

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

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

Phaedon Kyriakidis · Diofantos Hadjimitsis ·
Dimitrios Skarlatos · Ali Mansourian
Editors

Geospatial Technologies for Local and Regional Development

Proceedings of the 22nd AGILE Conference
on Geographic Information Science

 Springer

Editors

Phaedon Kyriakidis 
Department of Civil Engineering and
Geomatics
Cyprus University of Technology
Limassol, Cyprus

Dimitrios Skarlatos 
Department of Civil Engineering and
Geomatics
Cyprus University of Technology
Limassol, Cyprus

Diofantos Hadjimitsis 
Department of Civil Engineering and
Geomatics
Cyprus University of Technology
Limassol, Cyprus

Ali Mansourian 
Department of Physical Geography and
Ecosystem Science
Lund University GIS Center
Lund, Skåne Län, Sweden

ISSN 1863-2246

ISSN 1863-2351 (electronic)

Lecture Notes in Geoinformation and Cartography

ISBN 978-3-030-14744-0

ISBN 978-3-030-14745-7 (eBook)

<https://doi.org/10.1007/978-3-030-14745-7>

© Springer Nature Switzerland AG 2020

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

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

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

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

Organizing Committee

Scientific Programme Committee

Phaedon Kyriakidis, Cyprus University of Technology, Cyprus (Chair)
Diofantos Hadjimitsis, Cyprus University of Technology, Cyprus
Dimitrios Skarlatos, Cyprus University of Technology, Cyprus
Ali Mansourian, Lund University GIS Center, Sweden

Local Organizing Committee (at the Cyprus University of Technology, Cyprus)

Diofantos Hadjimitsis (Chair)
Dimitrios Skarlatos (Workshop Chair)
Phaedon Kyriakidis
Chris Danezis
Demetris Demetriou, Ministry of Transport, Communication and Works, Cyprus
Kyriacos Themistocleous
Athos Agapiou
Apostolos Papakonstantinou
Georgios Leventis
Stylianos Hadjipetrou
Ekaterini Kousoulou

Scientific Committee

Ana Paula Afonso, University of Lisbon, Portugal
Athos Agapiou, Cyprus University of Technology, Cyprus

Andreas Andreou, Cyprus University of Technology, Cyprus
Fernando Bacao, New University of Lisbon, Portugal
Marek Baranowski, Institute of Geodesy and Cartography, Poland
Melih Basaraner, Yildiz Technical University, Turkey
Giedre Beconyte, Vilnius University, Lithuania
Itzhak Benenson, Tel Aviv University, Israel
Lars Bernard, TU Dresden, Germany
Michela Bertolotto, University College Dublin, Ireland
Ralf Bill, Rostock University, Germany
Sandro Bimonte, IRSTEA, France
Thomas Blaschke, University of Salzburg, Austria
Lars Bodum, Aalborg University, Denmark
Arnold Bregt, Wageningen University, The Netherlands
Thomas Brinkhoff, Jade University Oldenburg, Germany
Dirk Burghardt, TU Dresden, Germany
Pedro Cabral, NOVA IMS, Portugal
Sven Casteleyn, University Jaume I, Spain
Christophe Claramunt, Naval Academy Research Council, France
Serena Coetzee, University of Pretoria, South Africa
Lex Comber, University of Leeds, UK
Joep Crompvoets, KU Leuven, Belgium
Çetin Cömert, Karadeniz Technical University, Turkey
Chris Danezis, Cyprus University of Technology, Cyprus
Sytze de Bruin, Wageningen University and Research, The Netherlands
Demetris Demetriou, Ministry of Transport, Communications and Works, Cyprus
Cécile Duchêne, IGN, France
Sara Irina Fabrikant, University of Zurich, Switzerland
Cidalia Fonte, University of Coimbra, Portugal
Jerome Gensel, University of Grenoble, France
Michael Gould, Esri and University Jaume I, Spain
Carlos Granell, University Jaume I, Spain
Diofantos Hadjimitsis, Cyprus University of Technology, Cyprus
Henning Sten Hansen, Aalborg University, Denmark
Lars Harrie, Lund University, Sweden
Jan-Henrik Haunert, University of Bonn, Germany
Roberto Henriques, Nova IMS, Portugal
Stephen Hirtle, University of Pittsburgh, USA
Hartwig Hochmair, University of Florida, USA
Joaquín Huerta, University Jaume I, Spain
Bashkim Idrizi, Geo-SEE Institute, Republic of Macedonia
Mike Jackson, University of Nottingham, UK
Bin Jiang, University of Gävle, Sweden
Derek Karssenbergh, Utrecht University, The Netherlands
Tomi Kauppinen, Aalto University, Finland
Marinos Kavouras, National Technical University of Athens, Greece

Dimitris Kotzinos, University of Cergy-Pontoise, France
Petr Kuba Kubicek, Masaryk University, Czech Republic
Phaedon Kyriakidis, Cyprus University of Technology, Cyprus
Patrick Laube, Zurich University of Applied Science, Switzerland
Francisco J Lopez-Pellicer, University of Zaragoza, Spain
Małgorzata Luc, Jagiellonian University, Poland
Ali Mansourian, Lund University GIS Center, Sweden
Bruno Martins, IST and INESC-ID, Portugal
Martijn Meijers, Delft University of Technology, The Netherlands
Filipe Meneses, University of Minho, Portugal
Peter Mooney, Maynooth University, Ireland
João Moura Pires, FCT/UNL, Portugal
Beniamino Murgante, University of Basilicata, Italy
Javier Nogueras-Iso, University of Zaragoza, Spain
Juha Oksanen, Finnish Geospatial Research Institute, Finland
Toshihiro Osaragi, Tokyo Institute of Technology, Japan
Ebba Ossiannilsson, Swedish Association for Distance Education, Sweden
Frank Ostermann, University of Twente, The Netherlands
Volker Paelke, Hochschule Ostwestfalen-Lippe, Germany
Marco Painho, New University of Lisbon, Portugal
Fragkiskos Papadopoulos, Cyprus University of Technology, Cyprus
Apostolos Papakonstantinou, Cyprus University of Technology, Cyprus
Petter Pilesjö, Lund University, Sweden
Poulicos Prastacos, FORTH, Greece
Hardy Pundt, HP, Germany
Ross Purves, University of Zurich, Switzerland
Martin Raubal, ETH Zürich, Switzerland
Wolfgang Reinhardt, UniBw Muenchen, Germany
Claus Rinner, Ryerson University, Canada
Jorge Rocha, University of Minho, Portugal
Armanda Rodrigues, New University of Lisbon, Portugal
Maribel Yasmina, Santos, University of Minho, Portugal
Tapani Sarjakoski, Finnish Geospatial Research Institute, Finland
Sven Schade, European Commission—DG JRC, Belgium
Christoph Schlieder, University of Bamberg, Germany
Monika Sester, Leibniz University of Hannover, Germany
Takeshi Shirabe, Royal Institute of Technology (KTH), Sweden
Dimitrios Skarlatos, Cyprus University of Technology, Cyprus
Jantien Stoter, Delft University of Technology, The Netherlands
Maguelonne Teisseire, UMR TETIS, France
Kyriacos Themistocleous, Cyprus University of Technology, Cyprus
Fred Toppen, Utrecht University, The Netherlands
Nico van de Weghe, Ghent University, Belgium
Ron van Lammeren, Wageningen University, The Netherlands
Jos van Orshoven, KU Leuven, Belgium

Danny Vandenbroucke, KU Leuven, Belgium
Lluís Vicens, University of Girona, Spain
Luis M. Vilches-Blázquez, Pontifical Xavierian University, Spain
Vít Voženílek, Palacky University Olomouc, Czech Republic
Monica Wachowicz, University of New Brunswick, Canada
Gudrun Wallentin, University of Salzburg, Austria
Robert Weibel, University of Zurich, Switzerland
Stephan Winter, The University of Melbourne, Australia
Bisheng Yang, Wuhan University, China
F. Javier Zarazaga-Soria, University of Zaragoza, Spain

Preface

AGILE, the Association of Geographic Information Laboratories in Europe (<https://agile-online.org>), aims to promote research and academic teaching on geographic information systems and science at the European level and beyond, as well as to stimulate and support networking activities between member laboratories (groups or departments). AGILE has been holding annual conferences at different locations across Europe since 1998, having varying themes and foci depending on current developments and trends in the field. The conferences attract a wide range of researchers and practitioners in geographic information science and allied fields, who develop and employ geospatial theories, methods and technologies for better spatial-enabled planning, decision-making and society engagement. AGILE annual conferences accept full papers, short papers, as well as poster contributions, and typically include a workshop day prior to the main conference. For more than a decade, full papers submitted to the AGILE annual conference have been published, after undergoing blind peer review, as a book by Springer International Publishing AG in its Lecture Notes in Geoinformation and Cartography series. It is this combination of breadth in thematic coverage and quality of research that has led to the annual AGILE conference been considered as the hallmark conference on geographic information science in Europe.

The 22nd AGILE conference was organized in collaboration with the Cyprus University of Technology, in Limassol, Cyprus, on 17–20 June 2019. The conference's central theme was *Geospatial Technologies for Local and Regional Development*, aiming to showcase the relevance of geospatial technologies, as well as the science behind those technologies, towards facilitating geoinformation-enabled innovation and sustainable economic growth not only in Europe but also in other countries of the Eastern Mediterranean region and the rest of the world. The contents of this volume comprise the 19 full papers that were selected by the Scientific Committee out of a set of 27 original submissions, and are divided into five parts. Part I consists of four chapters advancing the state of the art in *Geographic Information Representation, Retrieval and Visualization*, the basis upon which geospatial technologies are built. Part II, *Geoinformation Science and Geospatial Technologies in Transportation*, encompasses four chapters proposing

novel approaches for the analysis and modelling of spatial and/or spatiotemporal data in the field of transportation. Part III consists of six chapters putting forth new methods and applications of *Geoinformation Science and Geospatial Technologies in Urban/Regional Planning*. Part IV contains three chapters proposing novel analytical and modelling methodologies in apparently widely different fields, but nonetheless showcase the unifying role of *Spatial Scale as a Common Thread in Geoinformation Analysis and Modeling*. Lastly, Part V contains two important chapters exploring the *User and Workforce Dimensions of Geospatial Technologies*, two critical facets of the penetration and adoption of geospatial technologies worldwide. All in all, the breadth and depth of the contributions of this volume constitute a rather representative snapshot of the current developments in the field of geographic information science and technology, and help promote the role of *Geospatial Technologies for Local and Regional Development*.

Concluding, we would like to express our sincere gratitude to all the authors of the papers of this volume for their contributions, as well as the reviewers of the Scientific Committee for their respective evaluations. Both the quality of the contributions and the scrutiny of the evaluations were instrumental in maintaining the quality of the AGILE conference, as well as in enabling us to arrive at the final selection of the contents of this volume. We would also like to thank all those who helped editing this volume, as well as materializing the 22nd AGILE conference programme. We particularly thank the AGILE Council, which was instrumental in assisting with the scientific organization of the conference, as well as our colleagues at the Cyprus University of Technology for their enduring assistance towards making the conference a success. Lastly, we would like to express our gratitude to Springer International Publishing AG for their always helpful cooperation and assistance towards publishing and distributing this volume.

Limassol, Cyprus
Limassol, Cyprus
Limassol, Cyprus
Lund, Sweden
March 2019

Phaedon Kyriakidis
Diofantos Hadjimitsis
Dimitrios Skarlatos
Ali Mansourian

Contents

Part I Geographic Information Representation, Retrieval and Visualization	
Place Questions and Human-Generated Answers: A Data Analysis Approach 3	
Ehsan Hamzei, Haonan Li, Maria Vasardani, Timothy Baldwin, Stephan Winter and Martin Tomko	
Relaxing Unanswerable Geographic Questions Using A Spatially Explicit Knowledge Graph Embedding Model 21	
Gengchen Mai, Bo Yan, Krzysztof Janowicz and Rui Zhu	
Evaluating the Effectiveness of Embeddings in Representing the Structure of Geospatial Ontologies 41	
Federico Dassereto, Laura Di Rocco, Giovanna Guerrini and Michela Bertolotto	
Web-Based Visualization of Big Geospatial Vector Data 59	
Florian Zouhar and Julia Senner	
Part II Geoinformation Science and Geospatial Technologies in Transportation	
A Clustering-Based Framework for Understanding Individuals’ Travel Mode Choice Behavior 77	
Pengxiang Zhao, Dominik Bucher, Henry Martin and Martin Raubal	
Classification of Urban and Rural Routes Based on Motorcycle Riding Behaviour 95	
Gerhard Navratil, Ioannis Giannopoulos and Gilbert Kotzbek	
Route Choice Decisions of E-bike Users: Analysis of GPS Tracking Data in the Netherlands 109	
Gamze Dane, Tao Feng, Floor Luub and Theo Arentze	

Route Optimisation for Winter Maintenance 125
 Nikmal Raghestani and Carsten Keßler

**Part III Geoinformation Science and Geospatial Technologies
 in Urban/Regional Planning**

**Tracing Tourism Geographies with Google Trends:
 A Dutch Case Study** 145
 Andrea Ballatore, Simon Scheider and Bas Spierings

**Estimating the Spatial Distribution of Vacant Houses
 Using Public Municipal Data** 165
 Yuki Akiyama, Akihiro Ueda, Kenta Ouchi, Natsuki Ito, Yoshiya Ono,
 Hideo Takaoka and Kohta Hisadomi

**Enhancing the Use of Population Statistics Derived from Mobile
 Phone Users by Considering Building-Use Dependent
 Purpose of Stay** 185
 Toshihiro Osaragi and Ryo Kudo

**Potential of Crowdsourced Traces for Detecting Updates
 in Authoritative Geographic Data** 205
 Stefan S. Ivanovic, Ana-Maria Olteanu-Raimond, Sébastien Mustière
 and Thomas Devogele

**A Scalable Analytical Framework for Spatio-Temporal Analysis
 of Neighborhood Change: A Sequence Analysis Approach** 223
 Nikos Patias, Francisco Rowe and Stefano Cavazzi

**Improving Business-as-Usual Scenarios in Land Change
 Modelling by Extending the Calibration Period and Integrating
 Demographic Data** 243
 Romain Mejean, Martin Paegelow, Mehdi Saqalli and Doryan Kaced

**Part IV Spatial Scale as a Common Thread in Geoinformation
 Analysis and Modeling**

**Market Area Delineation for Airports to Predict the Spread
 of Infectious Disease** 263
 Carmen Huber and Claus Rinner

**Reflective Practice: Lessons Learnt by Using Board Games
 as a Design Tool for Location-Based Games** 291
 Catherine Jones and Konstantinos Papangelis

**Agent-Based Simulation for Indoor Manufacturing
 Environments—Evaluating the Effects of Spatialization** 309
 Stefan Kern and Johannes Scholz

Part V User and Workforce Dimensions of Geospatial Technologies

Towards a Usability Scale for Participatory GIS 327
Andrea Ballatore, Will McClintock, Grace Goldberg and Werner Kuhn

**Future Occupational Profiles in Earth Observation
and Geoinformation—Scenarios Resulting from Changing
Workflows** 349
Barbara Hofer, Stefan Lang and Nicole Ferber

Part I
**Geographic Information Representation,
Retrieval and Visualization**

Place Questions and Human-Generated Answers: A Data Analysis Approach



Ehsan Hamzei, Haonan Li, Maria Vasardani, Timothy Baldwin, Stephan Winter and Martin Tomko

Abstract This paper investigates place-related questions submitted to search systems and their human-generated answers. Place-based search is motivated by the need to identify places matching some criteria, to identify them in space or relative to other places, or to characterize the qualities of such places. Human place-related questions have thus far been insufficiently studied and differ strongly from typical keyword queries. They thus challenge today's search engines providing only rudimentary geographic information retrieval support. We undertake an analysis of the patterns in place-based questions using a large-scale dataset of questions/answers, MS MARCO V2.1. The results of this study reveal patterns that can inform the design of conversational search systems and in-situ assistance systems, such as autonomous vehicles.

Keywords Geographic information retrieval · Geographic questions · Question answering systems · Web search queries · Query classification

1 Introduction

In their everyday communication people frequently ask questions about places (Winter and Freksa 2012). This is reflected in the frequency of location-based queries in human-computer interaction, e.g., Web search (Sanderson and Kohler 2004) and question answering systems (Li and Roth 2006). The popularity of conversational bots and assistants (Radlinski and Craswell 2017) requires a shift from retrieving documents as response to keyword-based queries to natural answers to natural language questions.

A better understanding of the nuances expressed in search questions will enable better inference of the intent behind the query, and thus improve the delivery of tailored information in the answer. Due to polysemy of *place* references, their strong

E. Hamzei (✉) · H. Li · M. Vasardani · T. Baldwin · S. Winter · M. Tomko
The University of Melbourne, Parkville, Vic, Australia
e-mail: ehamzei@student.unimelb.edu.au

© Springer Nature Switzerland AG 2020
P. Kyriakidis et al. (eds.), *Geospatial Technologies for Local and Regional Development*, Lecture Notes in Geoinformation and Cartography,
https://doi.org/10.1007/978-3-030-14745-7_1

contextual dependence (Winter and Freksa 2012), diverse metaphorical uses (Agnew 2011), and complex, often vernacular references (Hollenstein and Purves 2010), the nuanced understanding and answering of place questions are still challenging for computer-based systems (Sui and Goodchild 2011). Place questions not only concern location (*where-questions*), but also encompass a wide range of informational needs about places from their types and affordances to their qualities. A thorough analysis of questions and answer sets is still missing, yet it is an essential prerequisite for future improvement of question interpretation and answer generation mechanisms.

In this paper, we investigate the structural patterns of place-related questions and their human-generated answers using the MS MARCO V2.1 dataset (Nguyen et al. 2016). By addressing our main research question: “*How place questions and their answers can be modelled based on place-related semantics?*”, we contribute:

- An analysis of the content of place-related questions and answers, by extracting patterns of place-related information through semantic encoding;
- A categorization of place-related questions and human-generated answers of the dataset based on semantic encoding and contextual semantic embedding;
- An investigation of the relations between the categories of questions and corresponding answers.

2 Literature Review

Research in geographic information retrieval from Web sources has focused primarily on geographic Web queries (Sanderson and Kohler 2004; Aloteibi and Sanderson 2014) and the geographic context of Web search (Backstrom et al. 2008). Geographical Web queries with *explicit* location (i.e., containing place names—*toponyms*) were analysed first by Sanderson and Kohler (2004). Later, implicit geographic queries relating to generic nouns designating locations known to a subgroup of people (e.g., a family sharing a location known as *home*) have also been identified (Wang et al. 2005).

In contrast, Jones et al. (2008) analyzed the distances between toponyms in the query and the location of the user indicated by their IP address to study the intent of the spatial search. They classified the distances into granular categories: same-city, same-state, same-country and different-country. Recently, Lee et al. (2015) used more accurate user positioning for a finer-grained analysis of search distances and correlated their distribution with the distribution of entities of different types in space. Yet, Web use patterns have since shifted dramatically towards mobile Web use (Lee et al. 2015), voice queries (Guy 2018), and conversational assistants (Radlinski and Craswell 2017). Voice queries are more similar to natural language questions than keyword queries (Guy 2018), and the transition from queries to questions and from result sets with links to automatically generated answers leads to a more natural interaction between humans and machines.

Place questions vary in focus, targeting a range of aspects of place (Wang et al. 2016). While two types of questions, ‘*where is*’, and ‘*how can I get to*’ are most common (Spink and Gunar 2001), people also ask questions to get information *about* place names, place types, and activities supported by the places themselves. While in human-human communications place questions are asked with different intent in mind such as gaining factual knowledge, recommendations, invitations, offers, and opinions (Wang et al. 2016), studies show that in human-computer interaction questions are mostly asked for informational and navigational purposes (Hollink 2008). ‘Where is’ questions aim at locating places in space, and thus understanding their spatial context. In a fundamental study, Shanon (1983) investigated how people answer such ‘where’ questions by extending the *room theory* to model human answering of ‘where’ questions. ‘How can I get to’ questions are a second dominant type of place questions, seeking procedural instructions to reach an intended destination (Tomko and Winter 2009). Answers to ‘where’ and ‘how can I get to’ questions carry descriptions of the environment in form of place or route descriptions, respectively (Winter et al. 2018). In both questions, the intended places can be referred to and described through place names, types, functions, and affordances. Yet, a detailed analysis of structural similarities and the patterns between the types of place questions and their answers has not been conducted thus far.

3 Data

We base our research on MS MARCO V2.1 (Nguyen et al. 2016). The dataset is a representative collection of questions (incl. place-related questions) and their human-generated answers. The current version of MS MARCO V2.1 contains more than one million records¹ of search question to Microsoft Bing, including human-generated answers, retrieved documents, and question types. The dataset identifies five types of questions—(1) LOCATION, (2) NUMERIC, (3) PERSON, (4) DESCRIPTION, and (5) ENTITY (Nguyen et al. 2016), but we focus exclusively on questions and answers which are labeled as LOCATION.

The original dataset has been divided into *train*, *dev*, and *test* sets. Here we use the *train* and *dev* subsets, with in total 56721 place-related questions. In MS MARCO V2.1,² the relation between questions and their answers is not necessarily one-to-one (Nguyen et al. 2016). Thus, we find questions with no answer, or more than one answer. Table 1 shows the number of answers per place question in the combined dataset, indicating that around 22% of questions have no answer and about 2% have multiple answers (i.e., due to question ambiguity or insufficient information in the retrieved documents).

¹<http://www.msmarco.org>.

²<https://github.com/dfcf93/MSMARCOV2/blob/master/README.md>.

Table 1 Number of answers per location question in the combined dataset

Number of answers (per question)	Count
0	12486
1	42884
2	1345
3	4
4	1
5	1

4 Method

We first translate place questions/answers into semantic encodings. Next, the questions and answers are clustered to identify common patterns in the data, based on semantic encodings and contrasted to a second approach based on embeddings. Finally, the relations between the questions and their answers are explored by linking the clusters of the questions and answers. In the following sections, we detail the method for generating the semantic encodings, categorizing the questions and answers, and analyzing the relation between questions and their answers.³

4.1 Semantic Encoding

We propose a semantic encoding schema extending that by Edwardes and Purves (2007). They introduced a mapping from part-of-speech to place semantics to extract elements, qualities, and affordances from generic nouns, adjectives, and verbs, respectively. We extend this semantic representation for the relatively short place questions and answers from MS MARCO, with the following primary elements of semantic encoding: (1) PLACE NAMES (e.g., *MIT*); (2) PLACE TYPES (e.g., *university*); (3) ACTIVITIES (e.g., *to study*); (4) SITUATIONS (e.g., *to live*); (5) QUALITATIVE SPATIAL RELATIONSHIPS (e.g., *near*); and (6) QUALITIES (e.g., *beautiful*). To differentiate types of questions, the WH- WORDS, and other generic OBJECTS are also considered.

Table 2 shows the resulting alphanumeric encoding schema. For example, after the removal of stop words, the question *what is the sunniest place in South Carolina* is translated into the encoding 2qtrn. This semantic encoding enables to analyze and categorize a large dataset of questions and answers by their structural patterns.

³The implementation is available at: <https://github.com/haonan-li/place-qa-AGILE19>.

Table 2 Semantic representation encoding

Semantic type	Part-of-speech	Code	Semantic type	Part-of-speech	Code
<i>where</i>	WH-word	1	Place name	Noun	n
<i>what</i>	WH-word	2	Place type	Noun	t
<i>which</i>	WH-word	3	Object	Noun	o
<i>when</i>	WH-word	4	Quality	Adjective	q
<i>how</i>	WH-word	5	Activity	Verb	a
<i>whom</i>	WH-word	6	Situation, and event	Verb	s
<i>whose</i>	WH-word	7	Spatial relationship	Preposition	r
<i>why</i>	WH-word	8			

4.2 Information Extraction

To extract further place-related semantics and linguistic information from the questions and their answers, we apply a sequence of preprocessing steps, based in part on the Stanford CoreNLP toolkit (Manning et al. 2014).

1. **Tokenization, tagging and dependency parsing:** First, the text is tokenized, and abbreviations are expanded into their canonical forms by using a common place name abbreviation table. Next, a part-of-speech tagger and dependency parser are applied to the text.
2. **Noun encoding:** Nouns are encoded in the order of place names (toponyms), place types, and objects. First, all subsequences of a sentence are considered candidate toponyms. These candidates are then matched with the GeoNames gazetteer⁴ to extract toponyms. We preference compound place names as better matches than simplex place names. Thus, *North Melbourne* is a compound place name and not an adjective followed by a place name. Next, nouns that are not place names have been considered as candidates for place type and objects. Place types are identified using dictionary lookup, and all nouns that are neither place names, nor place types are encoded as objects. The dictionary of generic place types is constructed by crawling the tag values of the OpenStreetMap (OSM) spatial database,⁵ due to the rich diversity of place types that are sourced from a large, global community of active volunteers.
3. **Verb encoding:** We have focused on verbs related to activities or situations (states) and ignored other types of verbs, known as accomplishments and developments (Mourelatos 1978). To differentiate between activities and situations, two sets of dynamic verbs (action and stative verbs) are collected and integrated from

⁴www.geonames.org.

⁵<https://www.openstreetmap.org>.

online resources.^{6, 7} Then, a pre-trained contextualized word embedding model ELMo (Peters et al. 2018), trained from large scale bidirectional language models, was used to extract activities and situations by considering contextual information. To compare the semantic similarity between the verbs in sentences with the verbs in the aforementioned verb sets, we first generate an embedding vector for each verb in the verb set. Then, for the extracted verbs from the questions and answers, their embedding vector representations are compared with the vectors in the two sets. Based on the computed Euclidean distances between the vector of the verb in a sentence with the vectors of the verbs in the two sets, the extracted verbs are classified as activities or situations.

4. **Preposition and adjective encoding:** Dependency parsing has been used to encode prepositions and adjectives. For the universal dependencies (De Marneffe et al. 2014) of each sentence, `case` and adjective modifier (`amod`) dependencies for prepositions and adjectives are extracted separately. Prepositions anchored to a place name are encoded as spatial relationships, and adjectives modifying a place type or a place name are considered as qualities of places.

4.3 Question/Answer Analysis

After extracting the desired place-related semantics and linguistic information, the questions and answers are translated into semantic encodings. To find the categories of place-related questions and answers, we compute clusters of the semantic representations. In this clustering, we first randomly select 1024 different encodings from the unique set of all semantic encodings, and subsequently re-model the semantic encodings of the questions/answers as 1024-dimensional vectors. The values in these vectors are calculated based on the Jaro similarity (Jaro 1989) between the semantic encodings of the questions/answers and the selected encodings. The Jaro similarity sim_j of two strings s_1, s_2 is given in Eq. 1:

$$sim_j = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3} \left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m} \right) & \text{otherwise} \end{cases} \quad (1)$$

where $|s_i|$ is the length of string s_i , m is the number of matching characters, and t is the number of transpositions. Specifically, two characters from two different strings are considered to match if they are the same and not farther than $\left\lfloor \frac{\max(|s_1|, |s_2|)}{2} \right\rfloor - 1$. The number of matching characters (but different sequence order) divided by 2, defines the number of transpositions.

Next, k -means clustering is applied to the questions and answers, separately. To measure whether the semantic representations retain the contextual similarity of the

⁶<https://www.gingersoftware.com/content/grammar-rules/verbs>.

⁷<https://www.perfect-english-grammar.com/support-files/stative-verbs-list.pdf>.

sentences, the ELMo representations for the questions and answers are used with the same similarity measure and clustering technique to provide a second set of clusters. The similarity of the clustering based on the semantic encodings and ELMo-based clusters is then evaluated. The results of clustering are also human-interpreted using the most frequent encodings in each cluster, enabling to derive the categories of place questions and answers.

In addition to categories of the questions and answers, we analyze their contents based on the frequently extracted place-related semantics. We also investigate the geographical distribution of the identified places using GeoNames to provide a deeper understanding of the dataset. In order to disambiguate between candidate toponyms, we combine the place names from the question, with those in the corresponding answers. The set of geospatial groundings that is associated with the minimum total distance between the toponyms is selected. In the case of a single toponym, this method cannot be applied. Consequently, only the resolved toponyms are used for describing the geographical distributions of the dataset. Finally, the relation between the questions with unresolved toponyms and unanswered questions is investigated.

4.4 Question/Answer Relationship

We link categories of questions to categories of answers to investigate the relationship between the content of place questions and their human-generated answers. As a category of place questions can be answered using one or more categories of answers (and vice-versa), we consider generic many-to-many relationships.

The same approach to categorization of questions and answers is also applied to concatenated question-answers. Finally, the result of linking categories of questions and categories of answers is compared with concatenated categories. For questions with multiple answers, multiple concatenated encodings are generated and investigated, while questions with no answers are concatenated with a unique pattern (○○).

5 Results

The results reveal three major groups of question/answers, replicated using both the semantic encoding and embedding approaches. We provide an initial qualitative interpretation of the patterns based on the encoding approach.

5.1 Preliminary Analysis

Place questions are constructed with a small number of tokens, and their answers are mostly short descriptions. Using tokenization and sentence segmentation, we find that 95.71% of questions contain less than ten tokens, and 98.17% of their answers are formulated with one, two, or three sentences.

5.1.1 Extracted Place-Related Semantics

Table 3 shows the top-five most frequent values extracted from the questions and answers for each type of place-related semantics. The dataset manifests a geographical bias to places in the USA. We identify similar patterns in geographical scales of place types and place names in the questions and their answers—i.e. related to coarse geographic scales, such as countries, and states. The frequency of activities and situations in the questions is notably higher than in the answers. In other words, people use activities and situations as criteria to describe the intended places in the questions rather than asking about activities and situations happening in a place – i.e., *where is the place with particular situation/affordance* is more often asked than *what is the affordance/situation of a specific place*. Thus, a set of detailed characteristics of a place may be specified in the question, leading to a simple answer with nominal references. Table 3 shows large differences between the frequencies of spatial relationships extracted from questions, and those extracted from answers. The reason is the use of spatial relationships for localization of places. Finally, the qualities included in questions and answers differ. In questions, qualities are used as criteria for identifying or describing the intended places, with most of the values being superlative adjectives. However, qualities in answers are mostly used to provide additional information about intended places or to describe particular places using combinations of quality and type—e.g., coastal region, or metropolitan area.

5.1.2 Toponym Resolution and Geographical Distribution

Tables 4 and 5 show that 79.2% and 68.1% of extracted places and question/answers records can be disambiguated, respectively. Figure 1 illustrates the geographical distribution of resolved places in the questions and answers. The spatial bias in the geographical distribution of places is visible. This spatial bias can be an artefact of the monolingual English dataset, population distribution of users of the source search engine (Microsoft Bing), or unknown question sampling from the overall Bing logs used to generate this dataset.

The comparison of resolved/unresolved and answered/unanswered records shows a strong relationship between ambiguous questions and unanswered questions. Table 5 shows the contingency table of this relation, revealing that 72.4% of unanswered questions are ambiguous (9042 records out of 12486 unanswered questions),

Table 3 Top-five frequent place-related semantics extracted from the dataset. Frequency in brackets

Type	In questions	In answers	Type	In questions	In answers
Place name	California (1393)	United States (4845)	Type	County (11702)	City (1714)
	Texas (1391)	California (1482)		State (2291)	State (1653)
	Florida (1148)	Texas (989)		City (1630)	County (1438)
	New York (895)	Florida (961)		Zone (745)	Area (882)
	Illinois (692)	New York (894)		Region (653)	Region (758)
Activity	Buy (340)	Go (64)	Situation	Find (1412)	Find (695)
	Go (296)	Run (62)		Live (746)	Have (405)
	Play (120)	Leave (55)		Have (662)	Live (305)
	Build (88)	Build (53)		Grow (321)	Include (231)
	See (86)	Move (38)		Originate (237)	Originate (125)
Spatial relation	In (3916)	In (10851)	Quality	Largest (242)	Largest (121)
	Near (153)	On (379)		Biggest (106)	Census-designated (68)
	At (142)	At (362)		Highest (97)	Metropolitan (54)
	On (109)	Near (275)		Expensive (56)	Small (46)
	Between (38)	Between (251)		Beautiful	Coastal (36)

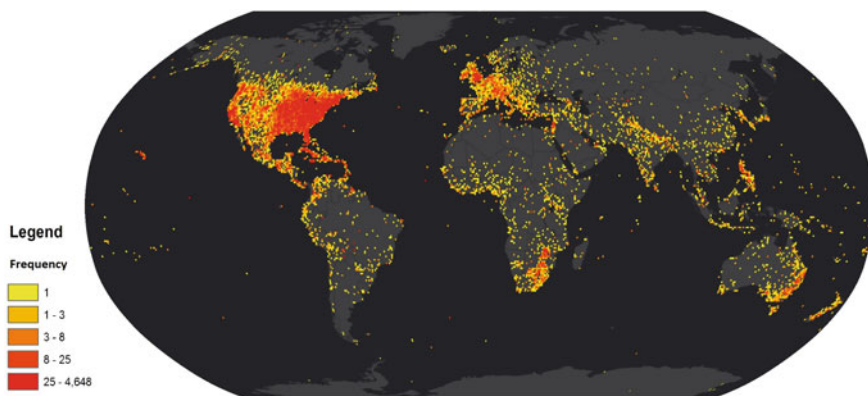
Table 4 Statistics of extraction, and disambiguation process

	Number of places	Percentage (%)
All	305868	100
Possible to resolve	242375	79.2
Ambiguous	63493	20.8

and 50.0% of ambiguous questions are not answered (9042 records out of 18093 ambiguous question/answer records). Hence, while people can interpret ambiguous toponyms considering other factors, such as saliency of places, and they can answer slightly more than half of the ambiguous questions, spatial ambiguity is one of the major reasons for unanswered questions. These questions may still be interpretable in context, yet this context is lacking in the MS MARCO dataset.

Table 5 Contingency table of ambiguous/resolved, and unanswered/answered questions

	Answered	Unanswered	Sum (percentage)
Resolved	35184	3444	38628 (68.1%)
Unresolved	9051	9042	18093 (31.9%)
Sum (percentage)	44.235 (78%)	12486 (22%)	56721 (100%)

**Fig. 1** Geographical distribution of resolved toponyms in questions and answers

5.2 Results of Question and Answer Analysis

We now investigate the structural patterns in place questions and answers separately, using the extracted place-related and linguistic semantic encoding. We identify distinct types of place question/answer pairs using a clustering approach based on the semantic representations with manual interpretation.

5.2.1 Frequent Patterns

Tables 6 and 7 present the top-five frequently observed semantic representation patterns. More than 40% of questions and answers can be described using the top-five representations. The top-five encodings for questions show that most of the spatial relationships are implicitly provided, e.g., *Barton County, Kansas* instead of *Barton County in Kansas*. In addition, the top-five questions are semantically different, first in terms of *What* and *Where* questions, and second in the manner the intended places are described. In questions with the semantic encoding $2\tau nn$ and $2\tau n$ the intended places are defined by type and spatial criteria, while questions with encoding $1n$ and $1nn$ use toponyms as verbal references to intended places. Finally, pattern $1o$, which is related to implicit situations (e.g., *where are orangutans* instead of *where*

Table 6 Frequent encoding patterns in questions

Pattern	Percentage (%)	Example
2tnn	12.5	<i>What is the county for Grand Forks North Dakota</i>
1n	8.6	<i>Where are the Boise Mountains</i>
1nn	7.8	<i>Where is Barton County Kansas</i>
1o	5.7	<i>Where are ores located</i>
2tn	5.5	<i>What is county for Seattle</i>

Table 7 Frequent encoding patterns in answers

Pattern	Percentage (%)	Example
n	26.2	<i>Germany</i> (question: <i>What country were gummy bears originally made in</i>)
o	6.2	<i>Hebrew</i> (question: <i>Where did letter j originate</i>)
nn	4.2	<i>Warren County, United States</i> (question: <i>Where is Lacona Iowa</i>)
rn	2.3	<i>In Worcester County</i> (question: <i>What county Fitchburg Massachusetts</i>)
nrnn	2.3	<i>Penhook is in Franklin County Virginia</i> (question: <i>What county is Penhook Virginia in</i>)

do orangutans live), or to non-geographical places (e.g., *where is the key button*), is frequently observed.

The most frequent patterns in answers provide evidence that people answer questions concisely. A single place name, which can be a compound or simple noun, is the answer for more than one quarter of the questions. We note that this may be an artifact of the interface used by people to answer the questions (presumably, keyboard). In addition, implicit relationships between places (e.g., nn as for *Tigard Oregon*) are also more frequent in answers than explicit spatial relationships. Interestingly, one of the frequent patterns of answers is related to non-geographical places, such as *Hebrew* as an answer for a *where* question about languages.

5.2.2 Categorizations of Place-Related Questions/Answers

To identify distinct types of place questions and answers, we used a clustering approach with manual interpretation based on the semantic representations of the results. Two different clustering techniques have been used for finding clusters of questions and answers based on word embeddings, and semantic encodings. We used the Calinski-Harabasz score (Caliński and Harabasz 1974) to find the optimum number of clusters for questions and answers, respectively. The score has been computed for all clusterings in the range of 2–30 clusters. For both clustering approaches and for both questions and answers, the score is optimal for three clusters. Importantly,

Table 8 Types of questions

ID	Type	Percentage (%)	Frequent patterns
1	Non-spatial <i>Place-related questions not aiming at localisation of places</i>	41.5	2tnn, 2tn, and 2tno
2	Spatial <i>Place-related questions about the location of places</i>	23.1	1n, 1nn, and 1nrn
3	Non-geographical, and ambiguous <i>Place-related questions about non-geographical places (e.g., fictional places) or ambiguous questions</i>	35.4	1o, 1os, and 1oo

Table 9 Types of answers

ID	Type	Percentage (%)	Frequent patterns
1	Explicit localization and spatial descriptions	25.8	nrnn, rnnn, and rnn
2	Implicit localization (place names, and addresses)	42.5	n, nn, and nnn
3	Non-geographical, and unanswered	31.7	oo, o, and ooo

we find that the results of clustering using semantic encoding and word embedding are highly similar, with a one-to-one relationship between the clusters with a 71.2% to 87.8% similarity using the Dice coefficient. To avoid presenting redundant information, here we report only the results of clustering based on semantic encoding which enables easier interpretability. The results of the clustering for questions and answers are interpreted manually using the most frequent semantic encodings in each cluster.

Tables 8 and 9 present the categories of questions/answers, and their overall percentage. We find that questions differ in terms of intention and formulation. While the first two types, non-spatial and spatial questions, relate to geographical places, the third type of questions is related to non-geographical entities such as virtual places (e.g., *Marvelous Bridge*, a place in the Pokemon World), and places in different types of space (e.g., *liver*, an organ in the space of the human body). In addition, non-spatial and spatial questions differ in the way the intended places are described in the questions. While in non-spatial questions places are described using place types, affordances, and situations, spatial questions contain toponyms as a direct reference to the intended places. In the first and second types of questions, spatial relationships to other places are frequently observed implicitly (e.g., 2nn, and 1nn), as well as explicitly (e.g., 1nrn, and 2trn).

The answers are also classified into three categories (Table 9). There is a noticeable difference between answer patterns relating to geographical and non-geographical places (the first two vs. the third category of answers). The difference between the

first two categories of answers derives from how the answer is formulated, using names and implicit relationships, or with spatial relationships which are explicitly mentioned in the answers. Moreover, the category of implicit localization includes notably fewer distinct semantic encodings—only 0.3% of all unique encodings, while the explicit localization category contains most of the patterns, 97% of all unique encodings. In other words, a broad range of simple to complex spatial descriptions are categorized as explicit localization.

5.3 Question/Answer Relationships

To investigate the relation between place questions and answers, two different scenarios are taken into account. First, we investigate the relation between types of questions and answers. Next, the concatenated encodings of questions and answers are investigated using a clustering approach and manual interpretation of the frequently observed patterns in each cluster.

5.3.1 From Questions to Answers

A many-to-many relationship between categories of questions and answers is considered and presented as a contingency table (Table 10). Spatial questions are mostly answered with implicit localization answers and are less ambiguous than non-spatial questions. This is because they are formulated with direct references to the intended places through toponyms. Non-spatial questions are more ambiguous and are mostly answered through complex descriptions or remain even unanswered. For example, *what are the best airports in Southern Utah?* is unanswered in the dataset, and it cannot be answered without describing what *best* means here. In addition, we observe relationships between non-geographical and ambiguous questions and implicit and explicit localization answers. Several reasons contribute to the observed relationship, such as human interpretation of ambiguous questions, issues in extraction of place-related semantics, which are propagated to the semantic encodings, and similarity of patterns in formulating questions and answers for geographical and non-geographical places impacting on clusterings. There is a strong relationship between non-spatial question and answers with explicit localizations. Three primary reasons for this are:

Table 10 Contingency table of linking the clusters

Q/A	Explicit localization	Implicit localization	Non-geographical
Non-spatial	9512	7966	6266
Spatial	2463	8400	2332
Non-geographical	2777	7970	9535

Table 11 Frequent patterns in concatenated clusters

ID	Frequent encoding	Example
1	2tnn-n, 2tn-n, and 2tnn-nrnn	<i>What county is Roselle NJ? Union County</i>
2	1n-n, 1nn-rnnn, and 1nn-n	<i>Where is Fairview? in Multnomah County Oregon United States</i>
3	1o-oo, 1oo-oo, and 1no-oo	<i>Where is the appendix to the liver? No Answer Present.</i>

1. **Answering style:** The content of question (how the intended places are defined) can also be repeated a part of answer (e.g., Question (2trnn): *What beach is closest to Busch Gardens Tampa?* Answer (nrnn): *Clearwater beach is the closest to Busch gardens, Tampa.*).
2. **Ambiguous questions:** Ambiguous questions are answered in more detail (partially) related to localization information (e.g., Question (2tn): *What city is Olongapo City?* Answer (nqtrnn): *Olongapo City is a 1st class highly urbanized city in Central Luzon Philippines.*).
3. **Misleading questions:** While the asker’s question is semantically a *where-question*, it is formulated as a *what-question* (e.g., Question (2nn): *What is Bentonville Arkansas County?* Answer (nrnn): *Bentonville is in Benton County Arkansas*). Here, the encoding captures semantics different from the intent of the asker and this therefore affects the categorization of the question.

5.3.2 Concatenated Clustering

In a last experiment, the semantic representations of questions and their corresponding answers are concatenated and then clustered. The results of clustering show (Table 11) that there is a direct one-to-one relationship between the categories of questions and the concatenated clusters. Using the Dice similarity coefficient, the similarity of first, second, and third clusters of questions and the first, second, and third clusters of questions concatenated with answers are 88.86%, 75.83%, and 75.98%, respectively.

6 Discussion and Conclusions

We analyzed the potential and limitations of MS MARCO V2.1—a large-scale dataset of place questions and human-generated answers—for place-related studies. We report on our approach using semantic encoding and clustering techniques to derive categories of place questions and answers.

The relations between place questions and answers are studied by linking categories of questions and answers, and concatenated clusters. We manually interpret

the most frequent patterns found to provide qualitative insights in place-querying behavior. Using geocoding and disambiguation techniques, we have successfully resolved the ambiguous place references in questions and answers. The geographical representation of places in questions and answers reveals that while the dataset has a global coverage, it is highly biased to North America, Western Europe, and Australia. While this bias may also be in part attributed to the biased coverage of the GeoNames gazetteer used for toponym resolution (Graham and De Sabbata 2015), users of the MS MARCO dataset are advised to be mindful of this coverage bias and its potential impact.

Through extraction of place-related semantics, we have found a semantic bias related to the scale of places. Analyzing the frequent place types shows that while questions to fine-grained places such as *schools*, *libraries*, and *airports* occur, most of the frequently asked place types are related to coarse geographic scales. The top-ten most frequent types of places identified in the questions and answers are associated with the *city to country level* scales. The diversity of activities, situations, qualities, and spatial relationships in the dataset shows further potential for studies in geographic information retrieval.

Expressing the patterns in questions and answers through the proposed semantic encoding proved to be useful for categorization. It produces results consistent with those achieved by the state-of-the-art word embedding approach using the pre-trained ELMo model. Yet, the encoding approach allows for manual inspection of the patterns and their qualitative interpretation, thus providing a valuable insight into place-related querying and answer-giving behavior.

Using a data mining approach and interpretation of the frequent patterns in the results, we found three main types of questions in MS MARCO V2.1—i.e., non-spatial, spatial, and non-geographical and ambiguous questions. Applying the same strategy to the answers similarly reveals three main categories of answers, explicit localization (spatial descriptions), implicit localization (names and addresses), and non-geographical and unanswered questions. By linking the questions and answers we show that even for spatial questions, people often answer with implicit localizations—e.g., *Where is Killaloe?* is answered with a single toponym, *Ireland*.

Our research reveals a high potential of the MS MARCO dataset for future place-related studies. It is, however, highly recommended to consider appropriate sampling strategies in order to deal with the geographical and semantic biases observed and discussed throughout this paper. Semantically richer representations that can differentiate more complex types of questions and answers may lead to a more nuanced characterization of the dataset and human place-querying behavior. Using sophisticated clustering approaches such *fuzzy C-means* may lead to derive better categorizations, however it remains as a future work of this study. Moreover, the proposed semantic encoding can be extended and used for other purposes such as translating place-related questions into computer-digestible queries.

Acknowledgements The support by the Australian Research Council grant DP170100109 is acknowledged.

References

- Agnew JA (2011) Space and place. In: Agnew J, Livingstone DN (eds) *The SAGE handbook of geographical knowledge*. SAGE Publications Ltd, London, pp 316–330
- Aloteibi S, Sanderson M (2014) Analyzing geographic query reformulation: an exploratory study. *J Am Soc Inf Sci Technol* 65(1):13–24
- Backstrom L, Kleinberg J, Kumar R, Novak J (2008) Spatial variation in search engine queries. In: *Proceedings of the 17th international conference on World Wide Web, WWW '08*, ACM, New York, USA, pp 357–366
- Calinski T, Harabasz J (1974) A dendrite method for cluster analysis. *Commun Stat* 3(1):1–27
- De Marneffe M-C, Dozat T, Silveira N, Haverinen K, Ginter F, Nivre J, Manning CD (2014) Universal stanford dependencies: a cross-linguistic typology. In: *Proceedings of the ninth international conference on language resources and evaluation*, vol 14, pp 4585–4592
- Edwardes AJ, Purves RS (2007) Eliciting concepts of place for text-based image retrieval. In: *Proceedings of the 4th ACM workshop on geographical information retrieval*, pp 15–18
- Graham M, De Sabbata S (2015) Mapping information wealth and poverty: the geography of gazetteers. *Environ Plan A* 47(6):1254–1264
- Guy I (2018) The characteristics of voice search: comparing spoken with typed-in mobile web search queries. *ACM Trans Inf Syst (TOIS)* 36(3):30:1–30:28
- Hollenstein L, Purves R (2010) Exploring place through user-generated content: using Flickr tags to describe city cores. *J Spat Inf Sci* 1(1):21–48
- Hollink JM V (2008) Effects of goal-oriented search suggestions. In: *BNAIC 2008 Belgian-Dutch conference on artificial intelligence*, p 177
- Jaro MA (1989) Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *J Am Stat Assoc* 84(406):414–420
- Jones R, Zhang WV, Rey B, Jhala P, Stipp E (2008) Geographic intention and modification in web search. *Int J Geogr Inf Sci* 22(3):229–246
- Lee C-J, Craswell N, Murdock V (2015) Inter-category variation in location search. In: *The 38th international ACM SIGIR conference on research and development in information retrieval*, ACM, 2767797, pp 863–866
- Li X, Roth D (2006) Learning question classifiers: the role of semantic information. *Nat Lang Eng* 12(3):229–249
- Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, McClosky D (2014) The Stanford CoreNLP natural language processing toolkit. In: *Association for computational linguistics (ACL) 2014 system demonstrations*, pp 55–60
- Mourelatos AP (1978) Events, processes, and states. *Linguist Philos* 2(3):415–434
- Nguyen T, Rosenberg M, Song X, Gao J, Tiwary S, Majumder R, Deng L (2016) MS MARCO: a human generated machine reading comprehension dataset. [arXiv:1611.09268](https://arxiv.org/abs/1611.09268)
- Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. In: *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies*, Volume 1 (Long Papers), pp 2227–2237
- Radlinski F, Craswell N (2017) A theoretical framework for conversational search. In: *Proceedings of the 2017 conference on conference human information interaction and retrieval*, pp 117–126
- Sanderson M, Kohler J (2004) Analyzing geographic queries. In: *SIGIR workshop on geographic information retrieval*, vol 2, pp 8–10
- Shanon B (1983) Answers to where-questions. *Discourse Process* 6(4):319–352
- Spink A, Gunar O (2001) E-commerce web queries: excite and ask jeeves study. *First Monday* 6(7)
- Sui D, Goodchild M (2011) The convergence of gis and social media: challenges for giscience. *Int J Geogr Inf Sci* 25(11):1737–1748
- Tomko M, Winter S (2009) Pragmatic construction of destination descriptions for urban environments. *Spat Cogn Comput* 9(1):1–29

- Wang L, Chen L, Dong M, Hussain I, Pan Z, Chen G (2016) Understanding user behavior of asking location-based questions on microblogs. *Int J Hum Comput Interact* 32(7):544–556
- Wang L, Wang C, Xie X, Forman J, Lu Y, Ma W-Y, Li Y (2005) Detecting dominant locations from search queries. In: *Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval*, pp 424–431
- Winter S, Freksa C (2012) Approaching the notion of place by contrast. *J Spat Inf Sci* 5(1):31–50
- Winter S, Hamzei E, Van De Weghe N, Ooms K (2018) *A graph representation for verbal indoor route descriptions. Spatial cognition*. Springer, Berlin

Relaxing Unanswerable Geographic Questions Using A Spatially Explicit Knowledge Graph Embedding Model



Gengchen Mai, Bo Yan, Krzysztof Janowicz and Rui Zhu

Abstract Recent years have witnessed a rapid increase in Question Answering (QA) research and products in both academic and industry. However, geographic question answering remained nearly untouched although geographic questions account for a substantial part of daily communication. Compared to general QA systems, geographic QA has its own uniqueness, one of which can be seen during the process of handling unanswerable questions. Since users typically focus on the geographic constraints when they ask questions, if the question is unanswerable based on the knowledge base used by a QA system, users should be provided with a relaxed query which takes distance decay into account during the query relaxation and rewriting process. In this work, we present a spatially explicit translational knowledge graph embedding model called TransGeo which utilizes an edge-weighted PageRank and sampling strategy to encode the distance decay into the embedding model training process. This embedding model is further applied to relax and rewrite unanswerable geographic questions. We carry out two evaluation tasks: link prediction as well as query relaxation/rewriting for an approximate answer prediction task. A geographic knowledge graph training/testing dataset, *DB18*, as well as an unanswerable geographic query dataset, *GeoUQ*, are constructed. Compared to four other baseline models, our TransGeo model shows substantial advantages in both tasks.

Keywords Geographic question answering · Query relaxation · Knowledge graph embedding · Spatially explicit model

G. Mai (✉) · B. Yan · K. Janowicz · R. Zhu
STKO Lab, UC Santa Barbara, Santa Barbara, CA 93106, USA
e-mail: gengchen_mai@geog.ucsb.edu

B. Yan
e-mail: boyan@geog.ucsb.edu

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

R. Zhu
e-mail: ruizhu@geog.ucsb.edu

1 Introduction

In the field of natural language processing, Question Answering (QA) refers to the methods, processes, and systems which allow users to ask questions in the form of natural language sentences and receive one or more answers, often in the form of sentences (Laurent et al. 2006). In the past decades, researchers from both academia and industry have been competing to provide better models for various subtasks of QA. Nowadays, many commercial QA systems are widely used in our daily life such as Apple Siri and Amazon Alexa.

Although QA systems have been studied and developed for a long time, geographic question answering remained nearly untouched. Although geographic questions account for a large part of the query sets in several QA datasets and are frequently used as illustrative examples (Yih et al. 2016; Liang et al. 2017), they are treated equally to other questions even though geographic questions are fundamentally different in several ways. First, many geographic questions are highly context-dependent and subjective. Although some geographic questions can be answered objectively and context independently such as *what is the location of the California Science Center*, the answers to many geographic questions vary according to when and where these questions are asked, and who asks them. Examples include *nightclubs near me that are 18+* (location-dependent), *how expensive is a ride from Stanford University to Googleplex* (time-dependent), and *how safe is Isla Vista* (subjective). Second, another characteristic of geographic questions is that the answers are typically derived from a sequence of spatial operations rather than extracted from a piece of unstructured text or retrieved from Knowledge Graphs (KG) which are the normal procedures for current QA systems. For example, the answer to the question *what is the shortest route from California Science Center to LAX* should be computed by a shortest path algorithm on a route dataset rather than searching in a text corpus. The third difference is that geographic questions are often affected by vagueness and uncertainty at the conceptual level (Bennett et al. 2008), thereby making questions such as *how many lakes are there in Michigan* difficult to answer.¹

Due to the previously mentioned reasons, it is likely to receive no answer given a geographic question. In the field of general QA such cases are handled by so-called (query) relaxation and rewriting techniques (Elbassuoni et al. 2011). We believe that geographic questions will benefit from *spatially-explicit relaxation methods* in which the spatial agency and time continuity should be taken into account during relaxation and rewriting. Interestingly, only a few researchers have been working on geographic question answering (Chen et al. 2013; Pulla et al. 2013; Scheider et al. 2018). In this paper, we will mainly focus on how to include spatial agency (distance decay effect) into the geographic query relaxation/rewriting framework.

The necessity of query relaxation/rewriting arises from the problem of *unanswerable questions* (Rajpurkar et al. 2018). Almost all QA systems answer a given question based on their internal knowledge bases (KB). According to the nature of such knowledge bases, current QA research can be classified into three categories:

¹Where the answer can vary between 63,000 and 10 depending on the conceptualization of Lake.

unstructured data-based QA (Rajpurkar et al. 2016; Miller et al. 2016; Yang et al. 2017; Chen et al. 2017; Mai et al. 2018), semi-structure table-based QA (Pasupat and Liang 2015), and structured-KB-based QA (so-called semantic parsing) (Yih et al. 2016; Liang et al. 2017, 2018; Berant et al. 2013). If the answer to a given question cannot be retrieved from these sources, this question will be called an *unanswerable question*. There are different reasons for unanswerable questions. The first reason is that the information this question focuses on is missing from the current KB. For example, if the question is *what is the weather like in Creston, California* (Question A) and if the weather information of Creston is missing in the current KB, the QA system will fail to answer it. Another reason may stem from logical inconsistencies of a given question. The question *which city spans Texas and Colorado* (Question B) is unanswerable no matter which KBs is used because these states are disjoint.

In order to handle these cases, the initial questions need to be relaxed or rewritten to answerable questions and spatial agency need to be considered in this process. A relaxed question to Question A can be *what is the weather like in San Luis Obispo County* because Creston is part of San Luis Obispo County. Another option is to rewrite Question A to a similar question: *what is the weather like in San Luis Obispo (City)* because San Luis Obispo is near to Creston. Which option to consider depends on the nature of the given geographic question. As for Question B, a relaxation solution would be to delete one of the contradictory conditions. Sensible query relaxation/rewriting should be based on both the similarity/relatedness among geographic entities (the distance decay effect) and the nature of the question. However, current relaxation/rewriting techniques (Elbassuoni et al. 2011; Fokou et al. 2017; Wang et al. 2018) do not consider spatial agency when handling unanswerable questions, and, thus, often return surprising and counter-intuitive results.

The research contributions of our work are as follows:

1. We propose a spatially explicit knowledge graph embedding model, TransGeo, which explicitly models the distance decay effect.
2. This spatially explicit embedding model is utilized to relax/rewrite unanswerable geographic queries. To the best of our knowledge, we are the first to consider the spatial agency between geographic entities in this process.
3. We present a benchmark dataset to evaluate the performance of the unanswerable geographic question handling framework. The evaluation results show that our spatially explicit embedding model outperforms non-spatial models.

The remainder of this work is structured as follows. In Sec. 2, several works about unanswerable question handling are discussed. Next, we present our spatially explicit KG embedding model, TransGeo, and show how to use this model to do unanswerable geographic question relaxation/rewriting in Sec. 3. Then, in Sec. 4 we empirically evaluate TransGeo against 4 other baseline models in two tasks: link predication task, unanswerable geographic question relaxation/rewriting and approximate answer prediction task. Then we conclude our work in Sec. 5 and point out the future research directions.

2 Related Work

The *unanswerable question* problem was recently prominently featured in the open domain question answering research field by Rajpurkar et al. (2018). The authors constructs a benchmark dataset, SQuADUn, by combining the existing Stanford Question Answering Dataset (SQuAD) with over 50,000 unanswerable questions. These new unanswerable questions are adversarially written by crowd-workers to *look similar to* the original answerable questions. In their paper, the unanswerable questions are used as negative samples to train a better QA model to discriminate unanswerable questions from answerable ones. In our work, we assume the question has already been parsed (e.g. to a SPARQL query) by a semantic parser and resulted in an empty answer set. The task is to relax or rewrite this question/SPARQL query and to generate a related query with its corresponding answer. In the Semantic Web field, SPARQL query relaxation aims to reformulate queries with too few or even no results such that the intention of the original query is preserved while a sufficient number of potential answers are generated (Elbassuoni et al. 2011).

Query relaxation models can be classified into four categories: similarity-based, rule-based, user-preference-based, and cooperative techniques-based models. Elbassuoni et al. (2011) proposed a similarity-based SPARQL query relaxation method by defining a similarity metric on entities in a knowledge graph. The similarity metric are defined based on a statistic language model over the context of entities. The relaxed queries are then generated and ranked based on this metric. This query relaxation method is defined purely based on the similarity between SPARQL queries. In contrast, our model jointly considers the similarity between queries and the probability that a selected answer to the *relaxed* query is, indeed, the answer to *original* query. This is possible due to the so-called Open World Assumption (OWA) commonly used by Web-scale KG by which statements/triples missing from the knowledge graph can still be true unless they are explicitly declared to be false within the knowledge graph. Our model aims at relaxing or rewriting a query such that the top ranked rewritten queries are more likely to generate the correct answer to the original one if it would be known.

With the increasing popularity of machine learning models in question answering and the Semantic Web, knowledge graph embedding models have been used to either predict answers for failed SPARQL queries (Hamilton et al. 2018) or recommend similar queries (Zhang et al. 2018; Wang et al. 2018). KG embedding models aim to learn distributional representations for components of a knowledge graph. Entities are usually represented as continuous vectors while relations, i.e., object properties, are typically represented as vectors (such as in TransE (Bordes et al. 2013), TransH (Wang et al. 2014), and TransRW (Mai et al. 2018)), matrices (e.g. TransR (Lin et al. 2015)), or tensors. For a comprehensive explanation of different KG embedding models, readers are referred to a recent survey by Wang et al. (2017).

Hamilton et al. (2018) proposes a graph query embeddings model (GQEs) to predict answers for conjunctive graph queries in incomplete knowledge graphs. GQEs first embeds graph nodes (entities) in a low-dimensional space and represents logical

operators as learned geometric operations (e.g., translation, rotation) in the embedding space. Based on the learned node embeddings and geometric operations, each conjunctive graph query can be converted into an embedding in a same embedding space. Then cosine similarity is used to compare the query embeddings and node embeddings, and subsequently rank the corresponding entities as potential answers to the current query. While GQEs have been successfully applied to representing conjunctive graph queries and entities in the same embedding space, they have some limitations. For instance, GQEs can only handle conjunctive graph queries, a subset of SPARQL queries. Additionally, the predicted answer to a conjunctive graph query is not associated with a relaxed/rewritten query as an explanation for the answer.

Wang et al. (2018) proposed an entity context preserving translational KG embedding model to represent each entity as a low-dimensional embedding and each predicate as a translation operation between entities. The authors show that compared with TransE (Bordes et al. 2013), the most popular and straightforward KG embedding model, their embedding model performs better in terms of approximating answers to empty answer SPARQL queries. They also present an algorithm to compute *similar* queries to the original SPARQL queries based on the approximated answers. Our work is developed based on this work by overcoming some limitations and including distance decay in the embedding model training process.

3 Method

Before introducing our spatially explicit KG embedding model, we briefly outline concepts relevant to our work.

Definition 1 Knowledge Graph: A knowledge graph (KG) is a data repository, which is typically organized as a directed multi-relational graph. Let $G = \langle E, R \rangle$ be a knowledge graph where E is a set of entities (nodes) and R is a set of relations (labeled edges). A triple $T_i = (h_i, r_i, t_i)$ can be interpreted as an edge connecting the head entity h_i (subject) with the tail entity t_i (object) by relation r_i (predicate).²

Definition 2 Entity Context: Given an entity $e \in E$ in the knowledge graph G , the context of e is defined as $C(e) = \{(r_c, e_c) | (e, r_c, e_c) \in G \vee (e_c, r_c, e) \in G\}$.

Definition 3 Basic Graph Pattern (BGP): Let V be a set of query variables in a SPARQL query (e.g., ?place). A basic graph pattern in a SPARQL query is a set of triple patterns (s_i, p_i, o_i) where $s_i, o_i \in E \cup V$ and $p_i \in R$. Put differently, we restrict triple patterns and thus BGP to cases where the variables are in the subject or object position.

²Note that in many knowledge graphs, a triple can include a datatype property as the relation where the tail is a literal. In our work, we do not consider these kind of triple as they are not used in any major current KG embedding model. We will use head (h), relation (r), and tail(t) when discussing embeddings and subject (s), predicate (p), object (o) when discussing Semantic Web knowledge graphs to stay in line with the literature from both fields.

Definition 4 SPARQL select query: For the purpose of this work, a SPARQL select³ query Q_j is defined as the form: $Q_j = \text{SELECT } V_j \text{ FROM } KG \text{ WHERE } GP$ where $V_j \subseteq V$ and KG is the studied knowledge graph and GP is a BGP.

The SPARQL query 1 shows an example which corresponds to the natural language question: *In which computer hardware company located in Cupertino is/was Steve Jobs a board member.* The answer should be `dbr:Apple_Inc`. If the triple (`dbr:Apple_Inc`, `dbo:locationCity`, `dbr:Cupertino,_California`), however, is missing from current KG, this question would become an unanswerable geographic question. Compared to the full SPARQL 1.1 language standard, two limitations of the given definition of a SPARQL query should be clarified:

1. Predicates in a SPARQL 1.1 BGP can also be a variables. Hence, Definition 3 presents a subset of all triple patterns, which can appear in a standard SPARQL query.
2. SPARQL 1.1 also contains other operations (UNION, OPTION, FILTER, LIMIT, etc.) not considered here and in related state-of-the-art work (Wang et al. 2018; Hamilton et al. 2018).

```

SELECT ?v
WHERE {
?v dbo:locationCity dbr:Cupertino,_California .
?v dbo:industry dbr:Computer_hardware .
dbr:Steve_Jobs dbo:board ?v .}

```

Listing 1 An example SPARQL query generated by a semantic parser

Given a SPARQL query Q_j parsed from a natural language geographic question, if executing Q_j on the current KG yields an empty answer set, our goal is: (1) learn a spatially explicit KG embedding model for the current KG which takes distance decay into account; (2) use the embedding model to infer a ranked list of approximated answers to this question; and (3) generate a relaxed/related SPARQL query for each approximate answer as an explanation for the query relaxation/rewriting process.

3.1 Modeling Geographic Entity Context in Knowledge Graphs

Based on the examples about relaxing or rewriting Question A and Question B in the introduction, we observe that a suitable query relaxation/rewriting for an unanswerable geographic question should consider both the similarity/relatedness among geographic entities (e.g., the distance decay effect) as well as the nature of the question. In terms of measuring semantic similarities among (geographic) entities in a

³We ignore ASK, CONSTRUCT, and DESCRIBE queries here as they are not typically used for question answering, and, thus, also not considered in related work.

knowledge graph, we borrow the assumption of distributional semantics from computational linguistic that *you shall know a word by the company it keeps* (Firth 1957). In analogy, the semantic similarity among (geographic) entities can be measured based on their contexts (Yan et al. 2017).

With regards to measuring the similarity/relatedness between general entities in a knowledge graph, both Elbassuoni et al. (2011) and Wang et al. (2018) consider the one degree neighborhood of the current entity as its context, which is shown in Definition 2. However, **this entity context modeling falls apart when geographic entities are considered in two ways**. First, this geographic entity context modeling does not fully reflect *Tobler's first law of geography*, which indicates that *near things are more related than distant things*. Since Definition 2 only considers object property triples as the entity context and disregard all datatype properties, all positional information, e.g., geographic coordinates, would not be considered in the context modeling. Although the place hierarchy is encoded as object property triples in most KG, e.g., GeoNames, GNIS-LD, and DBpedia, and these triples can also indirectly introduce distance decay effects into the context modeling, such contextual information is far too coarse. For example, Santa Barbara County, Los Angeles County, and Humboldt County are all subdivisions of California. From a place hierarchy perspective, all three should have the same relatedness to each other. But Santa Barbara County is more related to Los Angeles County rather than Humboldt County.

The second reason is due to the way geographic knowledge is represented in Web-scale knowledge graphs. For any given populated place, the place hierarchy of administrative units is modeled using the same canonical predicates. Put differently, even if no other triples are known about a small settlement, the KG will still contain at least a triple about a higher-order unit the place belongs to, e.g., a county. Consequently, all populated places in, say, Coconino County, Arizona, will share a common predicate (e.g., `dbo:isPartOf`) and object (e.g., `dbr:Coconino_County,_Arizona`). For tiny deserted settlements such as Two Guns, AZ this may also be the sole triple known about them. In contrast, major cities in the same county or state, e.g., Flagstaff, will only have a small percentage of their total object property triples be about geographic statements. This will result in places about which not much is known to have an artificially increased similarity.

These aforementioned two reasons demonstrate the necessity to model geographic entity context in a different way rather than Definition 2. In this work, we redefine Definition 2 by combing an edge-weighted PageRank and a sampling procedure. The underlying idea is to assign larger weights to geographic triples in an entity context where the weights are modeled from a distance decay function.

To provide a final and illustrative example of the problems that arise from embedding models that are not spatially explicit, consider the work by Wang et al. (2018). Their query example is *which actor is born in New York and starred in a United States drama film directed by Time Burton*. After passing the SPARQL version of this question to their query relaxation/rewriting model, the model suggests to change the birthplace from New York to Kentucky which is certainly a surprising relaxation from the original query. Although Kentucky is also a place as New York, it is too far

away from the birthplace, New York, the QA system user is interested in. A more reasonable relaxed/rewritten query should replace New York City with its nearby places, e.g. New Jersey.

3.2 Spatially Explicit KG Embedding Model

Given a knowledge graph $G = \langle E, R \rangle$, a set of geographic entities $P \subseteq E$, and a triple $T_i = (h_i, r_i, t_i) \in G$, we treat G as an undirected, unlabeled, edge-weighted multigraph MG , which means that we ignore the direction and label (predicate) for each triple in G . The weight $w(T_i)$ for triple T_i is defined in Eq. 1, where D is the longest (simplified) earth surface distance which is half of the length of the equator measured in kilometer; $dis(h_i, t_i)$ is the geodesic distance between geographic entity h_i and t_i on the surface of an ellipsoidal model of the earth measured in kilometer. The ε is a hyperparameter to handle the cases where h_i and t_i are collocated; and l is the lowest edge weight we allow for each triple. If the head place and tail place of a geographic triple are too far apart, we set its weight as the lower bound l , indicating that we do not expect strong spatial interaction at this distance.⁴

$$w(T_i) = \begin{cases} \max\left(\ln \frac{D}{dis(h_i, t_i) + \varepsilon}, l\right) & \text{if } h_i \in P \wedge t_i \in P \\ l & \text{otherwise} \end{cases} \quad (1)$$

The location of h_i and t_i are represented as their geographic coordinates stored in a knowledge graph, which are usually points. In this work, we use the `geo:geometry` property to get the coordinates of all geographic entities in DBpedia.

After we compute weights for each triple in MG , an edge-weighted PageRank is applied to this weighted multigraph, where edge weights are treated as the transition probability of the random walker from one entity node to its neighboring entity node. In order to prevent the random walker to get stuck at one *sinking node*, the PageRank algorithm also defines a teleport probability, which allows the random walker to jump to a random node in MG with a certain probability at each time step. Let $PR(e_i)$ be the PageRank score for each entity e_i in the knowledge graph, then $PR(e_i) \in (0, 1)$ represents the probability of a random walker to arrive at entity e_i after n time steps. If e_i had a lot of one degree triples (i.e., $|C(e_i)|$ is large), then e_i would have a larger $PR(e_i)$. Since $\sum_i PR(e_i) = 1$ and $|C(e_i)|$ have a long tail distribution, $PR(e_i)$ will also have a long tail distribution with few very large values but many small values. This skewed distribution would affect the later sampling process. In order to normalize $PR(e_i)$, we apply a *damping* function (Eq. 2). In Eq. 2, \ln is the natural log function; N is the number of entities in the knowledge graph G . This function has the nice

⁴We leave the fact that interaction depends on the travel mode and related issues for further work. Similarity, due to the nature of existing knowledge graphs, we use point data to represent places despite the problems this may introduce. Work on effectively integrating linestrings, polygons, and topology into Web-scale knowledge graphs is ongoing (Regalia et al. 2017).

property that $w(e_i)$ increases monotonically w.r.t. $PR(e_i)$ and the distribution of $w(e_i)$ is more normalized than $w(e_i)$. Therefore, $w(e_i)$ encodes the structural information of the original knowledge graph and the distance decay effect on interaction (and similarity/relatedness more broadly) among geographic entities. The more incoming and outgoing triples one entity e_i has, the larger its $w(e_i)$ will be. Also, the closer two geographic entities $e_i, e_j \in P$ are, the larger $w(e_i)$ and $w(e_j)$ would be.

$$w(e_i) = N \cdot \frac{\frac{1}{-\ln PR(e_i)}}{\sum_i \frac{1}{-\ln PR(e_i)}} \tag{2}$$

Next, we introduce the knowledge graph embedding model, which utilizes $w(e_i)$ as the distribution from which the entity context is sampled. Since $w(e_i)$ directly encodes the distance decay information among geographic entities, we call our model spatially explicit KG embedding model, denoted here as TransGeo.

Translation-based KG embedding models embed entities into low-dimensional vector spaces while relations are treated as translation operations in either the original embedding space (TransE) or relation-specific embedding space (TransH, TransR). This geometric interpretation provides us with a useful way to understand the embedding-based query relaxation/rewritten process.

Figure 1 shows the basic graph pattern of Query 1 and their vector representations in KG embedding space. If triple $(\text{dbr:Apple_Inc}, \text{dbo:locationCity}, \text{dbr:Cupertino}, _California)$ is missing from the current KG, this query becomes an unanswerable query. However, if we already obtained the learned embeddings for $e_1, e_2, e_3, r_1, r_2,$ and r_3 , we could compute the embedding of the query variable $?v$ with each triple pattern. Next, we can compute the weighted average of these embeddings to get the final embedding of $?v$, which is denoted as \mathbf{v} . Next, the k -nearest neighbor entities of \mathbf{v} can be obtained based on the cosine similarity between their embeddings. These k -nearest neighbor entities are treated as approximated answers to the original query 1. Based on each of these candidate answers,

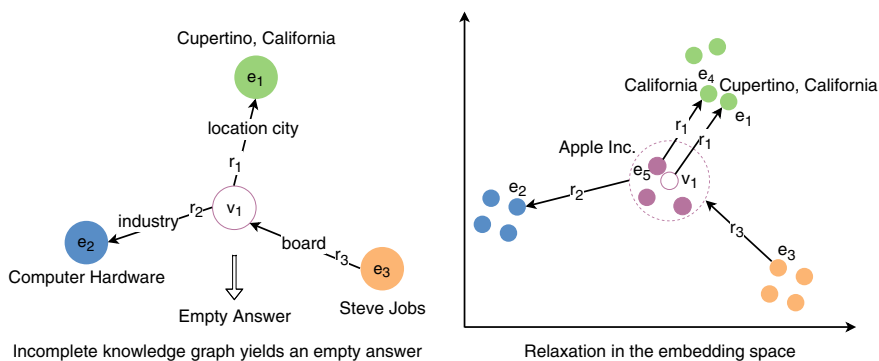


Fig. 1 An unanswerable geographic query example and its corresponding KG embedding

we cycle through each triple pattern in the original Query 1 to see whether they need to be relaxed or not, which is the major procedure for embedding-based query relaxation/rewritten.

In order to make the embedding-based query relaxation/rewriting process work well, the KG embedding model should be an entity context preserving model. However, one problem for the original TransE model is that each triple is treated independently in the training process which does not guarantee its context preservation. Inspired by the Continuous-Bag-of-word (CBOW) word embedding model (Mikolov et al. 2013), Wang et al. (2018) proposed an entity context preserved KG embedding model which predicts the *center* entity based on the entity context (Definition 2). However, as we discussed in Sec. 3.1, the geographic entity context can not be fully captured by using Definition 2 and we need another method to capture the distance decay effect, where $w(e_i)$ plays a role. Another shortcoming of the embedding model proposed in Wang et al. (2018) is that the size of entity context $|C(e_i)|$ varies among different entities which will make the number of triples trained in each batch different. This will have a negative effect on the model optimization process. Some entities may have thousands of incoming and outgoing triples, e.g., `dbr:United_States` has 232,573 context triples. This will imply that the model parameters will only update once all these triples are processed which is not a good optimization technique.

Based on this observation, we define a hyperparameter d as the context sampling size for each entity. If $|C(e_i)| > d$, then the context $C(e_i)$ of entity e_i would not be fully used in each KG embedding training step. Instead, the training context $C_{samp}(e_i)$ is sampled from $C(e_i)$ ($C_{samp}(e_i) \subseteq C(e_i)$) while the sampling probability of each context item (r_{ci}, e_{ci}) is calculated based on the damped PageRank value $w(e_{ci})$. If $|C(e_i)| > d$, the training context $C_{samp}(e_i)$ is sampled without replacement. If $|C(e_i)| < d$, $C_{samp}(e_i)$ is sampled with replacement. After a certain number of epochs t_{freq} , $C_{samp}(e_i)$ will be resampled for each entity. Because of this sampling strategy, a context item (r_{ci}, e_{ci}) of e_i would have a higher chance to be sampled if $e_i \in P \wedge e_{ci} \in P$, and e_i is close enough to e_{ci} in geographic space.

$$P(r_{ci}, e_{ci}) = \frac{w(e_{ci})}{\sum_{(r_{cj}, e_{cj}) \in C(e_i)} w(e_{cj})}, \text{ where } (e_i, r_{ci}, e_{ci}) \in G \vee (e_{ci}, r_{ci}, e_i) \in G \quad (3)$$

Based on the definition of entity training context $C_{samp}(e_i)$, a compatibility score between $C_{samp}(e_i)$ and an arbitrary entity e_k can be computed as Eq. 4, in which $\phi(e_k, r_{cj}, e_{cj})$ is the plausibility score function between (r_{cj}, e_{cj}) and e_k . In Eq. 5, $\|\cdot\|$ represents the L1-norm of the embedding vector; \mathbf{e}_k , \mathbf{e}_{cj} represent the KG embeddings for the corresponding entity e_k , e_{cj} , and \mathbf{r}_{cj} is the relation embedding of r_{cj} .

$$f(e_k, C_{samp}(e_i)) = \frac{1}{|C_{samp}(e_i)|} \cdot \sum_{(r_{cj}, e_{cj}) \in C_{samp}(e_i)} \phi(e_k, r_{cj}, e_{cj}) \quad (4)$$

$$\phi(e_k, r_{cj}, e_{cj}) = \begin{cases} \|\mathbf{e}_k + \mathbf{r}_{cj} - \mathbf{e}_{cj}\| & \text{if } (e_i, r_{cj}, e_{cj}) \in G \\ \|\mathbf{e}_{cj} + \mathbf{r}_{cj} - \mathbf{e}_k\| & \text{if } (e_{cj}, r_{cj}, e_i) \in G \end{cases} \quad (5)$$

The same assumption has been used here as TransE, which is that, in the *perfect* situation, if $(h_i, r_i, t_i) \in G$, $\|\mathbf{h}_i + \mathbf{r}_i - \mathbf{t}_i\| = 0$. Based on Eq. 4 and 5, if $e_k = e_i$, each $\phi(e_k, r_{cj}, e_{cj})$ would be small and close to zero, thus $f(e_i, C_{samp}(e_i))$ would be also small and close to zero. In contrast, if $C(e_k) \cap C(e_i) = \emptyset$, each $\phi(e_k, r_{cj}, e_{cj})$ would be very large and $f(e_i, C_{samp}(e_i))$ would also be large.

In order to set up the learning task, the pairwise ranking loss function has been used as the objective function like most KG embedding models do. Specifically, for each entity e_i , we randomly sample K entities as the negative sampling set $Neg(e_i)$ for e_i . Equation 6 shows the objective function of TransGeo, where γ is the margin and $max()$ is the maximum function.

$$\mathcal{L} = \sum_{e_i \in G} \sum_{e'_i \in Neg(e_i)} max(\gamma + f(e_i, C_{samp}(e_i)) - f(e'_i, C_{samp}(e_i)), 0) \quad (6)$$

3.3 KG Embedding Model Based Query Relaxation and Rewriting

After obtaining the learned TransGeomodel, we adopt the same procedure as Wang et al. (2018) to relax/rewrite the query. We briefly summarize the process below. We assume a SPARQL query Q with two variables $?v_1$ and $?v_2$, which are targets to be relaxed/rewritten in order to find approximated answers.

1. Given an empty answer SPARQL query Q , we partition the basic graph pattern into several groups such that all triple patterns in one group only contain one variable. Triples who have two variables $?v_1$ and $?v_2$ (connected triples) as its subject and object respectively are treated differently;
2. For each triple pattern group which contains variable $?v$, the embedding of $?v$ is first computed by each triple pattern based on the translation operations from the entity node to the variable node. Then the final embedding of $?v$ is computed as the weighted average of previous computed variable embeddings. The weight is calculated based on the number of matched triples of each triple patterns in the KG;
3. If Q has any connection triples, the embeddings of variables computed from each triple pattern group are refined based on the predicate of the connection edges. Then these embeddings will be treated as the final embeddings for each variable;
4. The approximate answers to each variable are determined by using their computed variable embeddings to search for the k -nearest embeddings of entities based on their cosine similarity. Each variable will have a ranked list of entities, e.g. $A(?v_1)$, $A(?v_2)$, as their approximated answers;

5. If Q has any connection triples, e.g. $(?v_1, r, ?v_2)$, we need to first use beam search to get top-K answer tuples for $?v_1$ and $?v_2$. And then each answer tuple (e_{1i}, e_{2j}) is checked for the condition $(e_{1i}, r, e_{2j}) \in G$. The answer tuples which satisfy this condition will be returned as a ranking list $Ans(Q)$ of approximated answers;
6. For each answer tuple $(e_{1i}, e_{2j}) \in Ans(Q)$, we enumerate each triple pattern to check the satisfaction. As for triple $(?v_1, r, e)$, if $(e_{1i}, r, e) \in G$, we do not perform any relaxation. If $(e_{1i}, r, e) \notin G$, then $(?v_1, r, e)$ will be relaxed based on Eq. 7. However, if e_{1i} does not have any outgoing triples, this triple pattern could not be relaxed and we would delete this triple pattern from the query relaxation/rewriting result. But the similarity score of this relaxation result will be set to 0;
7. The ranked list of answer tuples as well as the relaxed queries associated with them are returned to the users.

$$(e_{1i}, r_k, e_k) = \arg \max \left(\frac{\mathbf{r} \cdot \mathbf{r}_k}{\|\mathbf{r}\| \cdot \|\mathbf{r}_k\|} + \frac{\mathbf{e} \cdot \mathbf{e}_k}{\|\mathbf{e}\| \cdot \|\mathbf{e}_k\|} \right) \quad (7)$$

4 Experiment

Since almost all the established knowledge graph training dataset for KG embedding models, e.g., FB15K, WN18, do not contain enough geographic entities, we collect a new KG embedding training dataset, *DB18*,⁵ which is a subgraph of DBpedia. The dataset construction procedure is as follow: (1) We first selected all geographic entities which are part of `(dbo:isPartOf) dbr:California` with type `(rdf:type) dbo:City` which yields 462 geographic entities; (2) We use these entities as seeds to get their 1-degree and 2-degree object property triples and filter out triples with no `dbo:` properties; (3) we delete the entities and their associated triples whose node degree is less than 10; (4) we split the triple set into training and testing set and make sure that every entity and relation in the testing dataset will appear in training dataset. The statistic of *DB18* is listed in Table 1. ‘Geographic entities’ here means entities with a `geo:geometry` property.

Following the method we describe in Sec. 3.2, we compute the edge weights for each triple in *DB18* and an edge-weighted PageRank algorithm is applied on this undirected unlabeled multigraph. Here we set l to 1 and ε to 1. We select four models as the baseline models to compare with TransGeo: (1) *TransE*; (2) the context preserving translational KG embedding (Wang et al. 2018); (3) a simplified version of TransGeo in which the entity context items are randomly sampled from a uniform distribution, denoted as *TransGeo_{unweighted}*; (4) another simplified version of our model in which the PageRank are applied to unweighted multigraph, denoted as *TransGeo_{regular}*. We implement *TransE*, *TransGeo_{unweighted}*, *TransGeo_{regular}*, and TransGeo, in Tensorflow. We use the original Java implementation of Wang et al.

⁵<https://github.com/gengchenmai/TransGeo>.

Table 1 Summary statistic for *DB18*

DB18	Total	Training	Testing
# of triples	139155	138155	1000
# of entities	22061	–	–
# of relations	281	–	–
# of geographic entities	1681 (7.62%)	–	–

(2018).⁶ For all five models, we train them for 1000 epochs with the margin $\gamma = 1.0$ and learning rate $\alpha = 0.001$. As for *TransGeo_{unweighted}*, *TransGeo_{regular}*, and *TransGeo*, we use 30 as the entity context sampling size d and 1000 as batch size. We resample the entity context every 100 epochs. As for the context preserving translational KG embedding (Wang et al. 2018), we use 10 as the entity context size cut-off value. The embedding dimension of all these five embedding models is 50.

In order to demonstrate the effectiveness of our spatially explicit KG embedding model, *TransGeo*, over the other four baseline models, we evaluate these five KG embedding models in two task: the standard link predication task and an relaxation/rewriting task to predict answers to the otherwise unanswerable geographic questions. The evaluation results are listed in Table 2.

The common link prediction task is used to validate the translation preserving characteristic of different models. The set up of the link prediction task follows the evaluation protocol of Bordes et al. (2013). Given a correct triple $T_k = (h_k, r_k, t_k)$ from the testing dataset of DB18, we replace the head entity h_k (or tail t_k) with all other entities from the dictionary of DB18. The plausibility scores for each of those n triples are computed based on the plausibility score functions of *TransE* ($\| \mathbf{h} + \mathbf{r} - \mathbf{t} \parallel$). Then these triples are ranked in ascending order according to this score. The higher the correct triple ranks in this list, the better this learned model. Note that some of the corrupted triples may also appear in the KG. For example, as for triple (*dbr:Santa_Barbara, _California, dbo:isPartOf, dbr:California*), if we replace the head *dbr:Santa_Barbara, _California* with *dbr:San_Francisco*, the result corrupted triple (*dbr:San_Francisco, dbo:isPartOf, dbr:California*) is still in the DBpedia KG. These false negative samples need to be filtered out. Mean reciprocal rank (*MRR*) and *HIT@10* are used as evaluation metrics where *Raw* and *Filter* indicate the evaluation results on the original ranking of triples or the filtered list which filters out the false negative samples. According to Table 2, *TransGeo* performs the best in most of the metrics and the only metric *TransGeo* cannot outperform is *MRR* in the raw setting. This evaluation shows that our spatially explicit model does indeed hold the translation preserving characteristic.

For the quality of the unanswerable geographic query relaxation/rewriting results, we evaluate the results based on the ranking of the approximate answers (Hamilton et al. 2018). Let’s take Question 1 as an example. One reason which causes an empty

⁶<https://github.com/wangmengsd/re>.

Table 2 Two evaluation tasks for different KG embedding models

	Link prediction				SPARQL relaxation	
	MRR		HIT@10		MRR	HIT@10
	Raw	Filter	Raw (%)	Filter (%)		
<i>TransE</i> model	0.122	0.149	30.00	34.00	0.008	5% (1 out of 20)
Wang et al. (2018)	0.113	0.154	27.20	30.50	0.000	0% (0 out of 20)
<i>TransGeo_{regular}</i>	0.094	0.129	28.50	33.40	0.098	25% (5 out of 20)
<i>TransGeo_{unweighted}</i>	0.108	0.152	30.80	37.80	0.043	15% (3 out of 20)
TransGeo	0.104	0.159	32.40	42.10	0.109	30% (6 out of 20)

answer is that some triples were missing from the KG, e.g., (`dbr:Apple_Inc`, `dbo:locationCity`, `dbr:Cupertino,_California`), and the current SPARQL query is overly restrictive. However, based on the KG embedding model, we can approximate the embeddings of the variables in the current query. This variable embeddings can be used to search for the most *probable* answers/entities to each variable in the embedding space. These *k-nearest* entities are assumed to be more probable to be the correct answer of the original question. The correct answer (based on the Open World Assumption) to Question 1 is `dbr:Apple_Inc`. If the KG embedding is good at preserving the context of entities, the embedding of `dbr:Apple_Inc` will appear close to the computed variable embedding (See Fig. 1). So the performance of the query relaxation/rewriting algorithm can be evaluated by checking the rank of the correct answer in the returned ranking list of the approximate answers.

Based on the above discussion, we construct another evaluation dataset, *GeoUQ*, which is composed of 20 unanswerable geographic questions. Let G_{train} be a knowledge graph which is composed of all the training triples of DB18⁷ and G_{all} be a knowledge graph containing all training and testing triples in DB18.⁸ Both G_{train} and G_{all} can be accessed through the SPARQL endpoint. These queries satisfy 2 conditions: 1) each query Q will yield empty answer set when executing Q on G_{train} ; 2) Q will return only one answer when executing Q on G_{all} . The reason for making Q a one-answer query in G_{all} is that the user also expects one answer from the QA system to the question (s)he poses. One-answer queries are also the common setup for many QA benchmark datasets, e.g. WikiMovie (Miller et al. 2016), WebQuestionsSP (Yih et al. 2016). *MRR* and *HIT@10* are used as evaluation metrics for this task.

All five KG embedding models are evaluated based on the same query relaxation/rewriting implementation. The evaluation results are shown in Table 2. From Table 2, we can conclude that TransGeo outperform all the other baselines models both on *MRR* and *HIT@10*.

⁷<http://stko-testing.geog.ucsb.edu:3080/dataset.html?tab=query&ds=/GeoQA-Train>.

⁸<http://stko-testing.geog.ucsb.edu:3080/dataset.html?tab=query&ds=/GeoQA-All>.

Table 3 Query relaxation/rewriting results of different KG embedding models for Query 1

	Relaxation 1	Relaxation 2	Relaxation 3
<i>TransE</i>	<p>Query: SELECT ?v WHERE { ?v db:locationCity dbr:Fountain_Valley_California . ?v db:industry dbr:Data_storage_device . dbr:John_Tu db:knownFor ?v . } Answer: dbr:Kingston_Technology</p>	<p>Query: SELECT ?v WHERE { ?v db:locationCity dbr:Agoura_Hills_California . ?v db:industry dbr:Interactive_entertainment . dbr:Heavy_Iron_Studios db:owningCompany ?v . } Answer: dbr:THQ</p>	<p>Query: SELECT ?v WHERE { ?v db:locationCity dbr:Pasadena_California . ?v db:industry dbr:Entertainment . } Answer: dbr:Landmark_Entertainment_Group</p>
Wang et al. (2018)	<p>Query: SELECT ?v WHERE { ?v db:locationCity dbr:Irvine_California . ?v db:industry dbr:Video_game_industry . dbr:Activision_Blizzard db:division ?v . } Answer: dbr:Blizzard_Entertainment</p>	<p>Query: SELECT ?v WHERE { ?v db:locationCity dbr:Mountain_View_California . ?v db:industry dbr:Interactive_entertainment . } Answer: dbr:Paragon_Studios</p>	<p>Query: SELECT ?v WHERE { ?v db:locationCity dbr:San_Mateo_California . ?v db:industry dbr:Video_game . } Answer: dbr:Digital_Pictures</p>
<i>TransGeoUnweighted</i>	<p>Query: SELECT ?v WHERE { ?v db:locationCity dbr:Palo_Alto_California . ?v db:industry dbr:Computer_hardware . dbr:William_Redington_Hewlett db:knownFor ?v . } Answer: dbr:Hewlett-Packard</p>	<p>Query: SELECT ?v WHERE { ?v db:locationCity dbr:California . ?v db:keyPerson dbr:Elon_Musk . dbr:Elon_Musk db:knownFor ?v . } Answer: dbr:SolarCity</p>	<p>Query: SELECT ?v WHERE { ?v db:locationCity dbr:Santa_Clara_California . ?v db:industry dbr:Computer_hardware . } Answer: dbr:Console_Inc</p>

(continued)

Table 3 (continued)

	Relaxation 1	Relaxation 2	Relaxation 3
<i>TransGeoregular</i>	<p>Query: SELECT ?v WHERE { ?v dbo:foundationPlace dbr:Santa_Clara,_California . ?v dbo:industry dbr:Computer_hardware . } Answer: dbr:Iasomi_Networks</p>	<p>Query: SELECT ?v WHERE { ?v dbo:location dbr:San_Carlos,_California . ?v dbo:industry dbr:Computer_hardware . } Answer: dbr:Check_Point</p>	<p>Query: SELECT ?v WHERE { ?v dbo:locationCity dbr:Lake_Forest,_California . ?v dbo:industry dbr:Computer_hardware . } Answer: dbr:PSSC_Labs</p>
<i>TransGeo</i>	<p>Query: SELECT ?v WHERE { ?v dbo:locationCity dbr:Redwood_City,_California . ?v dbo:industry dbr:Computer_hardware . dbr:Steve_Jobs dbo:occupation ?v . } Answer: dbr:NeXT</p>	<p>Query: SELECT ?v WHERE { ?v dbo:locationCity dbr:California . ?v dbo:industry dbr:Computer_hardware . dbr:Steve_Jobs dbo:board ?v . } Answer: dbr:Apple_Inc</p>	<p>Query: SELECT ?v WHERE { ?v dbo:foundationPlace dbr:Stoux_City,_Iowa . ?v dbo:industry dbr:Computer_hardware . dbr:EMachines dbo:owningCompany ?v . } Answer: dbr:Gateway_Inc</p>

Table 3 show the top 3 query relaxation/rewriting results of Question 1 from all the 5 KG embedding models. For each query, the highlighted part in the BGP is the part where the query is changed from the original Query 1. Note that some of the relaxation/rewriting results have less triple patterns than the original Query 1. This is because the current approximate answer/entity does not have any outgoing or incoming triples to be set as the alternative to the original triple pattern. Hence, we delete this triple pattern. This has been described in Step 6 in Sec. 3.3. From Table 3, we can see that the correct answer `dbr:Apple_Inc` has been listed as the second approximate answer for TransGeo. However, all the 4 baseline models fail to predict this correct answer in their top 10 approximate answers list. Besides the perspective of predicting the correct answers, we can also evaluate the models by inspecting the quality of the relaxed/rewritten queries. For example, the top 1 relaxed query from TransGeo changes `dbr:Cupertino,_California` to `dbr:Redwood_City,_California` which is a nearby city of `dbr:Cupertino,_California`. Although the predicted answer is `dbr:NeXT` rather than `dbe:Apple_Inc`, this query relaxation/rewriting makes sense and is meaningful for the user. The 2nd relaxed result from TransGeo changes `dbr:Cupertino,_California` to `dbr:California` which is a superdivision of `dbr:Cupertino,_California`. This is indeed a real *query relaxation* which relaxes the geographic constraint to its superdivision. In short, our spatially explicit KG embedding model, TransGeo, produces better result than all baseline models.

5 Conclusion

In this work, we discussed why geographic question answering differs from general QA in general, and what this implies for relaxation and rewriting of empty queries specifically. We demonstrated why distance decay has to be included explicitly in the training of knowledge graph embeddings and showed cases of neglecting to do so. As a result, we propose a spatially explicit KG embedding models, TransGeo, which utilizes an edge-weighted PageRank and sampling strategy to include the distance decay effect into the KG embedding model training. We constructed a geographic knowledge graph training dataset, *DB18* and evaluated TransGeo as well as four baseline models. We also created an unanswerable geographic question dataset (*GeoUQ*) for two evaluation tasks: link prediction and answer prediction by relaxation/rewriting. Empirical experiments show that our spatially explicit embedding model, TransGeo, can outperform all the other 4 baseline methods on both task. As for the link prediction task, in the filter setting, our model outperforms the other baselines by at least 3.2% at *MRR* and 11.4% at *HIT@10*. In terms of the unanswerable geographic question approximate answer prediction task, our model outperform the other 4 baselines by at least 11.2% at *MRR* and 20% at *HIT@10*.

In terms of future work, firstly, the distance decay information is explicitly encoded into our KG embedding model which gives up on flexibility, e.g., to model modes of transportation. In the future, we want to explore ways to only consider distance

decay during query relaxation rather than the model training step. Secondly, as for the method to compute the edge weights of the knowledge graph, we used point geometries which may yield misleading results for larger geographic areas such as states. This limitation is due to the availability of existing knowledge graphs. Work to support more complex geometries and topology is under way.

References

- Bennett B, Mullenby D, Third A (2008) An ontology for grounding vague geographic terms. In: FOIS, vol 183, pp 280–293
- Berant J, Chou A, Frostig R, Liang P (2013) Semantic parsing on freebase from question-answer pairs. In: Proceedings of the 2013 conference on empirical methods in natural language processing, pp 1533–1544
- Bordes A, Usunier N, Garcia-Duran A, Weston J, Yakhnenko O (2013) Translating embeddings for modeling multi-relational data. *Adv Neural Inf Process Syst* 27:2787–2795
- Chen D, Fisch A, Weston J, Bordes A (2017) Reading wikipedia to answer open-domain questions. In: Proceedings of the 55th annual meeting of the association for computational linguistics (Volume 1: long papers), vol 1, pp 1870–1879
- Chen W, Fosler-Lussier E, Xiao N, Raje S, Ramnath R, Sui D (2013) A synergistic framework for geographic question answering. In: 2013 IEEE seventh international conference on semantic computing (ICSC), IEEE, pp 94–99
- Elbassuoni S, Ramanath M, Weikum G (2011) Query relaxation for entity-relationship search. In: Extended semantic web conference. Springer, Berlin, pp 62–76
- Firth JR (1957) A synopsis of linguistic theory, 1930–1955. *Studies linguist Anal*
- Fokou G, Jean S, Hadjali A, Baron M (2017) Handling failing rdf queries: from diagnosis to relaxation. *Knowl Inf Syst* 50(1):167–195
- Hamilton W, Bajaj P, Zitnik M, Jurafsky D, Leskovec J (2018) Embedding logical queries on knowledge graphs. *Adv Neural Inf Process Syst* 2027–2038
- Laurent D, Séguéla P, Nègre S (2006) QA better than IR?. In: Proceedings of the workshop on multilingual question answering. Association for Computational Linguistics, pp 1–8
- Liang C, Berant J, Le Q, Forbus KD, Lao N (2017) Neural symbolic machines: learning semantic parsers on freebase with weak supervision. In: Proceedings of the 55th annual meeting of the association for computational linguistics (Volume 1: long papers), vol 1, pp 23–33
- Liang C, Norouzi M, Berant J, Le QV, Lao N (2018) Memory augmented policy optimization for program synthesis and semantic parsing. *Adv Neural Inf Process Syst* 10014–10026
- Lin Y, Liu Z, Sun M, Liu Y, Zhu X (2015) Learning entity and relation embeddings for knowledge graph completion. *AAAI* 15:2181–2187
- Mai G, Janowicz K, He C, Liu S, Lao N (2018) POIReviewQA: a semantically enriched POI retrieval and question answering dataset. In: Proceedings of the 12th workshop on geographic information retrieval, ACM, p 5
- Mai G, Janowicz K, Yan B (2018) Support and centrality: learning weights for knowledge graph embedding models. In: International conference on knowledge engineering and knowledge management. Springer, Berlin, pp 212–227
- Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. *Adv Neural Inf Process Syst* 31:1111–1119
- Miller A, Fisch A, Dodge J, Karimi A-H, Bordes A, Weston J (2016) Key-value memory networks for directly reading documents. In: Empirical methods in natural language processing (EMNLP), pp 1400–1409
- Pasupat P, Liang P (2015) Compositional semantic parsing on semi-structured tables. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th

- international joint conference on natural language processing (Volume 1: Long Papers), vol 1, pp 1470–1480
- Pulla VS, Jammi CS, Tiwari P, Gjoka M, Markopoulou A (2013) Questcrowd: a location-based question answering system with participation incentives. In: 2013 IEEE conference on computer communications workshops (INFOCOM WKSHPs), IEEE, pp 75–76
- Rajpurkar P, Jia R, Liang P (2018) Know what you don't know: unanswerable questions for SQuAD. [arXiv:1806.03822](https://arxiv.org/abs/1806.03822)
- Rajpurkar P, Zhang J, Lopyrev K, Liang P (2016) SQuAD: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 conference on empirical methods in natural language processing, pp 2383–2392
- Regalia B, Janowicz K, McKenzie G (2017) Revisiting the representation of and need for raw geometries on the linked data web. In: LDOW@ WWW
- Scheider S, Ballatore A, Lemmens R (2018) Finding and sharing GIS methods based on the questions they answer. *Int J Digit Earth* 1–20
- Wang M, Wang R, Liu J, Chen Y, Zhang L, Qi G (2018) Towards empty answers in sparql: Approximating querying with rdf embedding. In: International semantic web conference. Springer, Berlin, pp 513–529
- Wang Q, Mao Z, Wang B, Guo L (2017) Knowledge graph embedding: a survey of approaches and applications. *IEEE Trans Knowl Data Eng* 29(12):2724–2743
- Wang Z, Zhang J, Feng J, Chen Z (2014) Knowledge graph embedding by translating on hyperplanes. *AAAI* 14:1112–1119
- Yan B, Janowicz K, Mai G, Gao S (2017) From itdl to place2vec: reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts. In: Proceedings of the 25th ACM SIGSPATIAL international conference on advances in geographic information systems, ACM, p 35
- Yang F, Nie J, Cohen WW, Lao N (2017) Learning to organize knowledge with n-gram machines. [arXiv:1711.06744](https://arxiv.org/abs/1711.06744)
- Yih W-t, Richardson M, Meek C, Chang M-W, Suh J (2016) The value of semantic parse labeling for knowledge base question answering. In: Proceedings of the 54th annual meeting of the association for computational linguistics (Volume 2: short papers), vol 2, pp 201–206
- Zhang L, Zhang X, Feng Z (2018) TrQuery: an embedding-based framework for recommending sparql queries. [arXiv:1806.06205](https://arxiv.org/abs/1806.06205)

Evaluating the Effectiveness of Embeddings in Representing the Structure of Geospatial Ontologies



Federico Dassereto, Laura Di Rocco, Giovanna Guerrini
and Michela Bertolotto

Abstract Nowadays word embeddings are used for many natural language processing (NLP) tasks thanks to their ability of capturing the semantic relations between words. Word embeddings have been mostly used to solve traditional NLP problems, such as question answering, textual entailment and sentiment analysis. This work proposes a new way of thinking about word embeddings that exploits them in order to represent geographical knowledge (e.g., geographical ontologies). We also propose metrics for evaluating the effectiveness of an embedding with respect to the ontological structure on which it is created both in an absolute way and with reference to its application within geolocation algorithms.

Keywords Geospatial ontologies · Word embeddings · NLP · Embedding evaluation · Geo-term similarity

1 Introduction

Huge amounts of data are easily produced at increasing velocity, favored by the growth of social networks and the increasing internet coverage. Thanks to the nature of social media, these data tend to encapsulate information in short texts: these kinds of data are called microblogs. Microblog analysis can reveal interesting knowledge about social dynamics and the propagation of information. Specifically, by analyzing

F. Dassereto (✉) · L. Di Rocco · G. Guerrini
Università di Genova, DIBRIS, Via Dodecaneso 35, Genova, Italy
e-mail: federico.dassereto@edu.unige.it

L. Di Rocco
e-mail: laura.dirocco@dibris.unige.it

G. Guerrini
e-mail: giovanna.guerrini@unige.it

M. Bertolotto
School of Computer Science, University College Dublin, Belfield, Dublin 4, Ireland
e-mail: michela.bertolotto@ucd.ie

the semantic content of a microblog, one might be able to understand what the user wanted to describe and where the message originated from.

The semantic content of a message could be sufficient to reach a good accuracy in the geolocation process at the city level of detail, when machine learning techniques (like classification) are applied (Dredze et al. 2016). However, for certain applications city level detail is not enough. To reach a good sub-city level accuracy some external knowledge can be added to further exploit semantics. The external knowledge is supposed to encapsulate the hierarchical nature of the geographic area the geolocation process is working on. One of the advantages of using an external knowledge is that a prior training is not required, opening the door to real-time applications.

Word embeddings are exploited in many natural language processing tasks where words or phrases from the vocabulary are mapped to vectors of real numbers. Conceptually, they are a mathematical embedding from a space with one dimension per word to a continuous vector space with a much lower dimension (Bengio et al. 2003). Recently, word embeddings emerged as a new way of encapsulating hierarchical information (Nickel and Kiela 2017). These approaches are linked to the introduction of hyperbolic embeddings, which are able to capture semantic correlation between objects, representing them in a continuous way. Evaluating the effectiveness of an embedding with respect to the structure that it generates is not a simple task, since a standard methodology or metric does not exist. Traditional embedding evaluation processes do not evaluate them in an absolute way, but assess the effectiveness of an embedding only with respect to a specific task, as we will discuss in Sect. 2. Decoupling the evaluation of an embedding from a specific task can be an interesting step towards understanding both the meaning of the dimensions of an embedding (obscure in literature) and the correlation between human and machine perception of semantic closeness. This particular gap is a real problem when one needs to build an ad-hoc embedding, because its quality clearly affects the accuracy of the result.

The main contributions of this work are three-fold: (i) presenting word embeddings as an alternative representation for geographical ontologies; (ii) providing a task-independent metric for evaluating the effectiveness of such structures with respect to the original ontology structure; (iii) performing an extensive experimental evaluation of the embedding effectiveness for a specific case study both from a task-dependent and from a task-independent point of view. Contribution (ii) represents the first attempt to evaluate an embedding independently of the task for which it is created, and represents the most important and ambitious part of this work. The proposed evaluation abstracts from the task, it just focuses on the inner structure of the ontologies. The experimental evaluation (contribution (iii)) aims at assessing the coherence of the task-independent metrics with the results achieved with reference to a specific task. In general, the model we propose is suitable for every kind of hierarchical structure. Our case study focuses on geographical ontologies.

The remainder of this paper is organized as follows: Sect. 2 discusses preliminaries and related work on geographical ontologies and embeddings. Section 3 defines distance between geo-terms, on ontologies and embeddings, and the notion of distortion, on which the task-independent embedding evaluation is based. Section 4

introduces the geolocation algorithm that we use for the task-dependent evaluation. The performed experiments and the obtained results are described in Sect. 5. Finally, conclusions and future work directions are drawn in Sect. 6.

2 Preliminaries and Related Work

In this section, we discuss the background of our work and related approaches. We start introducing the concepts of ontologies and embeddings, and then we outline the state of the art on word embeddings, with a focus on current trends on embedding evaluation. In the paper, we consider a specific use case for embeddings: geolocation algorithms. Since geolocation algorithms are not the focus of this paper, for further information we refer the readers to the survey (Zheng et al. 2018).

2.1 Geographical Ontologies

An ontology is defined as “a specification of a conceptualization” (Gruber et al. 1993). Ontologies are useful tools for representing the semantic scope of a domain, by defining concepts and relations, and sharing knowledge among users. Ontologies have been used for a number of tasks: improving communication among agents (human or software), reusing data models, developing knowledge schemas, etc. All these tasks deal with interoperability issues and can refer to different domains. A geographical ontology is an ontology that describes a set of geographical entities in a hierarchical structure.

A geographical ontology is composed of a set of objects (e.g. roads, hotels) and relations between them. We can define an ontology as a set of pairs (*top*, *class*) where *top* represents an object and *class* represents one of the classes to which the object belongs; the relation between them could mean that *top is_subclass_of class* or *top is_a instance of class*. We obtain a graph $G = (V, E)$ where V is the set of toponyms and classes and E is the set of relations between them. In such a system of classes and subclasses, we define *instance node* a node on the left hands side of the *is_a* relation. Notice that, such graph is a metric space. Therefore, a distance function defines this space.

Many ontologies have been proposed in the literature. Examples include the FAO Geopolitical Ontology¹ which tracks historical changes from 1985 up until today and the DBpedia² ontology which has been manually created based on the most commonly used infoboxes within Wikipedia and, therefore, contains also geographic information.

¹<https://www.fao.org/countryprofiles/geoinfo/en/>.

²<https://wiki.dbpedia.org/>.

GeoNames³ is a geosemantic data source available and accessible through various web services. The GeoNames database contains over 10 millions geographical names corresponding to over 7.5 millions unique features. GeoNames consists of various locations and is divided into nine feature classes, and then subcategorized in one of the 645 subclasses. Each of the nine classes has many subclasses. A specific node in the Geonames ontology could be linked to different classes or subclasses, because different objects could share the same name.

Stadler et al. (2012) presented an ontology called LinkedGeoData (LGD) that links OSM information to DBpedia, GeoNames and others ontologies. LGD uses the comprehensive OpenStreetMap⁴ spatial data collection to create a large spatial knowledge base. It consists of more than 3 billion nodes and 300 million ways and the resulting RDF data comprises approximately 20 billion triples. The data are available according to the Linked Data principles and interlinked with DBpedia and GeoNames.

2.2 Word Embeddings

Since we consider geographical data from a non-geometric point of view, we need a mathematical structure that encapsulates the structure of our ontologies. The need for this structure emerges because textual data are not easily understandable by computers, although it is trivial for humans to understand similarities and relations.

Embeddings. An embedding is a mapping from a metric space O to another metric space S :

$$f : O \rightarrow S \quad (1)$$

A metric space is an ordered pair (M, d) where M is a set and d is a distance function:

$$d : M \times M \rightarrow \mathbb{R} \quad (2)$$

The most common use of embeddings is mapping discrete objects (like words) to vectors of real numbers. The main advantages of embedding discrete objects in a vector space are the dimensionality reduction (while maintaining the relations between the elements) and the possibility of easily defining similarity functions between elements (exploiting the vectors operations). Wang et al. (2017) presented an extensive survey on geospatial data embedding (especially knowledge graph). There are different ways to build an embedding starting from a text or a hierarchical structure: a common way is to train a neural network on a text (Mikolov et al. 2013; Jia et al. 2014) or by projecting the data into a particular geometry model like the Poincaré disk model (Nickel and Kiela 2017). Other embedding methods for geospatial data exists (Kejriwal and Szekely 2017) but the authors do not show the usefulness of

³<https://www.geonames.org>.

⁴<https://www.openstreetmap.org>.

this embedding neither for geolocation algorithms nor for learning task. An interesting family of embeddings is that of hyperbolic embeddings. Such embeddings lie in hyperbolic spaces, which are suitable to represent the hierarchical structure and maintain distances among elements (Krioukov et al. 2010). The idea behind hyperbolic embeddings is very simple: forcing elements with semantic correlations to be closest to each other in the embedding space (hyperbolic). Many hyperbolic embeddings based on different models of hyperbolic geometry have been proposed. Examples include the Poincaré Disk Model (Nickel and Kiela 2017; Sala et al. 2018) and the Lorentz Model (Nickel and Kiela 2018).

Word embeddings. The idea on which word embeddings rely is that languages have a distributional structure (Harris 1954). Word embeddings are numerical representations of texts where the dimensionality is reduced and the relations between words are maintained. The intuitions on which word embeddings are built are the following:

- similar words appear in similar contexts⁵;
- word vectors are computed taking into account contexts;
- similar words have similar vectors.

The last point of the previous listing is the key to popularity of word embedding: the similarity between words is reduced to similarity between vectors.

There are two ways to obtain word embeddings:

- Load into your model word embeddings in conjunction with different machine learning tasks and apply them for the task that you are trying to solve. This kind of embedding is called *pre-trained word embedding*.
- Learn word embeddings jointly with the main task at hand, e.g., document classification or sentiment analysis.

In the first case, we are able to embed a specific text (or any other structure) in a low-dimensional vector space, within which we can choose the embedding parameters and decide to emphasize some kind of relations. On the other hand, pre-trained word embeddings are produced exploiting huge knowledge and computational power. The most famous pre-trained word embeddings are *Word2vec* (Mikolov et al. 2013) and *GloVe* (Pennington et al. 2014).

In the following we will refer to the space where words are mapped as *semantic space*.

2.3 *Embeddings Evaluation*

Griffiths et al. (2007) first introduced the concept of evaluation for embeddings; by developing *Word2Vec* (Mikolov et al. 2013) proposed a novel approach of evaluation. The first complete evaluation on word count and prediction (Baroni et al. 2014) shows

⁵The context is understood as a fixed window around a target word.

that the prediction models like Word2Vec outperform the counting models. Recently, Bojanowski et al. (2017) improved the Word2vec model by boosting the computation time and making possible to represent words not seen during the training phase. They also proposed improvements for translation through vector models. In 2017, Bakarov published an exhaustive survey on evaluation methods (Bakarov 2018).

At the First Workshop on Evaluating Vector Space Representations for NLP, Faruqui et al. (2016) presented a set of problems associated with word similarity evaluation of word vector models:

- *Obscureness of the notion of semantics*: refers to the fact that we consider “as good” an embedding that reflects our understanding of semantics, but we cannot ensure whether our understanding is absolutely correct.
- *Lack of proper training data*: the problem is the absence of data split into training and testing sets, so researchers adjust the parameter for their data.
- *Absence of correlation between intrinsic and extrinsic methods*: refers to the lack of correlation between the two main categories of evaluation.
- *Lack of significance tests*: relates to the absence of “benchmarks”.

Extrinsic evaluation is based on the ability of word embeddings to be used as the feature vectors of supervised machine learning algorithms (like Maximum Entropy Model) e.g., the downstream NLP tasks. Nayak et al. proposed a set of tasks to test the embedding on Nayak et al. (2016), including Semantic Role Labelling, Text Classification and Part-of-Speech Tagging. Schnabel et al. (2015) demonstrated that extrinsic evaluation for word embeddings trained for serving in a wide range of different tasks fails when there is no correlation between tasks. The current idea about extrinsic evaluation is that it cannot be used as a general evaluation model, but just for highlighting some useful properties of certain embeddings. Intrinsic evaluations are experiments in which word embeddings are compared with human judgments on words relations. The main intrinsic evaluation models are (i) Word semantic similarity (Baroni et al. 2014), where similar words are supposed to have similar vectors; (ii) Word analogy (Sayeed et al. 2016), which is based on the idea that arithmetic operations in a word vector space could be predicted by humans, and (iii) the Thematic fit (Sayeed et al. 2016), where the idea is to assess how well word embeddings could find most semantically similar nouns for a certain verb that is used in a certain role. All these evaluation models are built to give a measure of how suitable a word embedding is to play a certain role for a specific task. No measure has been proposed to evaluate how “good” an embedding is with respect to the original text (or structure) on which it was generated.

3 Task-Independent Metrics

In this section, we propose a task-independent metric for assessing the quality of our embeddings. We first discuss how to measure the distance between geo-terms, i.e.,

terms that refer to objects in the physical world that have a dual nature (semantic and spatial), which are associated as labels with nodes in a geographical ontology, then introduce how distances are measured in the target space and finally we define our metric which evaluates the embeddings in terms of distortion, i.e., how much we expand or contract the distances in the embedding process.

3.1 Distance in the Ontology

Intuitively, the distance between geo-terms should capture the distance of corresponding nodes in the graph. Considering the shortest path between nodes introduces semantic noise in the measurements, because such path could pass through other instance nodes. Moreover, the shortest path cannot capture all the semantic meanings of a node, it just represents closeness in terms of edges. In order to capture semantic closeness among nodes, what we want is that the path from a node to another one encapsulates the classes and subclasses chain, emphasizing the hierarchical structure. This constraint is enforced also by the nature of hyperbolic embeddings to well represent complex hierarchical structures. Thus, *in a path only the departure and arrival nodes can be leaves, all others nodes must be classes or subclasses.*

To model this constraint, we introduce function $\sigma(x)$ that produces a list in which the different semantic meanings of a term are split into different objects. It associates each node in the graph with a set of nodes in a tree. Given a graph $G = (V, E)$ and an instance node $x \in V$, a mapping function σ is thus a function such that:

$$\sigma(x) = [x_0, \dots, x_{deg(x)}] \quad (3)$$

where $deg(x)$ represents the degree of node x in the ontology graph.

We notice that, for the nodes in the graph with edges representing the *is_subclass_of* relation in the ontology, σ is simply the identity function. By applying function σ to all the nodes in the ontology graph G (or equivalently only to the instance nodes) we obtain a tree that represents the ontology with edges representing different semantic meanings of a term. Figure 1 shows the differences between the graph and the tree obtained by applying σ .

Once applied the conversion from graph to tree, we need to redefine the distance function on the ontology.

Definition 1 (*Distance function in a graph*) Given a graph $G = (V, E)$, its tree T representation and two elements $a, b \in V$, the *average distance function* between a and b is defined as:

$$adf(a, b) = \frac{\sum_{a' \in \sigma(a)} \sum_{b' \in \sigma(b)} SP(a', b')}{\#(\sigma(a)) \cdot \#(\sigma(b))}$$

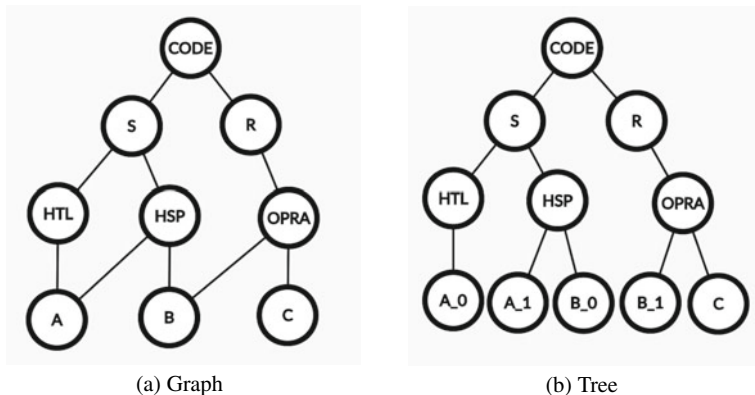


Fig. 1 Graph to tree conversion

where $SP(a', b')$ represents the shortest path in the tree (which is also the only possible simple path) and $\#(\cdot)$ represents the length of the semantic meanings list (i.e. the number of occurrences of \cdot in G).

Notice that, since a graph can be defined as a metric space, we need a metric. The adf is a metric on G .

Proof Given $x, y, z \in G$ the following conditions are satisfied:

- *Non-negativity*: $adf(x, y) \geq 0$ is satisfied since G is a non-weighted graph.
- *Identity*: $adf(x, y) = 0 \iff x = y$ is satisfied by the definition of SP distance.
- *Symmetry*: $adf(x, y) = adf(y, x)$ is satisfied since G is a non-directed graph. As we are interested in easily computable distances, we consider the graph as undirected in order to define the shortest path (SP) and the average distance function (adf).
- *Triangle inequality*: $adf(x, y) \leq adf(x, z) + adf(z, y)$ is satisfied since

$$\begin{aligned}
 adf(x, z) + adf(z, y) &= \frac{1}{\#(\sigma(x)) \cdot \#(\sigma(z))} \sum_{x' \in \sigma(x)} \sum_{z' \in \sigma(z)} SP(x', z') + \\
 &\frac{1}{\#(\sigma(z)) \cdot \#(\sigma(y))} \sum_{z' \in \sigma(z)} \sum_{y' \in \sigma(y)} SP(z', y') = \\
 &\frac{1}{[\#(\sigma(x)) \cdot \#(\sigma(z))] [\#(\sigma(z)) \cdot \#(\sigma(y))]} \left[[\#(\sigma(z)) \cdot \#(\sigma(y))] \sum_{x' \in \sigma(x)} \sum_{z' \in \sigma(z)} SP(x', z') + \right. \\
 &\left. [\#(\sigma(x)) \cdot \#(\sigma(z))] \sum_{z' \in \sigma(z)} \sum_{y' \in \sigma(y)} SP(z', y') \right] = \\
 &\frac{1}{\#(\sigma(x)) \cdot \#(\sigma(y)) \cdot \#(\sigma(z))^2} \#(\sigma(z)) \left[\sum_{y' \in \sigma(y)} \sum_{x' \in \sigma(x)} \sum_{z' \in \sigma(z)} SP(x', z') + \right. \\
 &\left. \sum_{x' \in \sigma(x)} \sum_{y' \in \sigma(y)} \sum_{z' \in \sigma(z)} SP(z', y') \right] = \\
 &\frac{1}{\#(\sigma(x)) \cdot \#(\sigma(y)) \cdot \#(\sigma(z))} \sum_{z' \in \sigma(z)} \sum_{x' \in \sigma(x)} \sum_{y' \in \sigma(y)} SP(x', z') + SP(z', y') \geq
 \end{aligned}$$

$$\begin{aligned}
& [SP(x', z') + SP(z', y')] \geq SP(x', y') \rightarrow \\
& \frac{1}{\#(\sigma(x)) \cdot \#(\sigma(y)) \cdot \#(\sigma(z))} \sum_{x' \in \sigma(x)} \sum_{y' \in \sigma(y)} SP(x', y') = \\
& \frac{\#(\sigma(x)) \cdot \#(\sigma(y))}{\#(\sigma(x)) \cdot \#(\sigma(y))} adf(x, y)
\end{aligned}$$

Hence, $adf(x, y)$ is a metric. \square

3.2 Distance in the Hyperbolic Embedding

Since we are dealing with hyperbolic embeddings in the Poincaré Disk Model (Nickel and Kiela 2017), the hyperbolic distance function for such space is defined as:

$$d_H(x, y) = acosh \left(1 + 2 \frac{\|x - y\|^2}{(1 - \|x\|^2)(1 - \|y\|^2)} \right) \quad (4)$$

where x, y are points in the model and $\|\cdot\|$ is the Euclidean norm.

This equation has a strong correlation with hierarchical structures like ontologies. As the Euclidean norm of the two points increases, the arc of circumference that connects them tends to pass through the origin of the disk. In ontologies (and in general in hierarchical structures) the shortest path between elements is the path which passes through the parents of such elements. This motivates the power of hyperbolic embedding in representing hierarchical structures.

3.3 Evaluation Metric: Distortion

The evaluation metric must provide a global analysis of the embedding and the ontology, so we need to consider all the possible pairs of elements. The simplest and most efficient way to design an evaluation metric is based on distances. The metric evaluates the embeddings in terms of *distortion*: it calculates how much we expand or contract the distances in the embedding process (Sala et al. 2018). We choose to evaluate the embedding in terms of distortion because it is the concept that best fits the idea of representation, since our goal is to propose embeddings as ontologies representation. Generally, values close to 0 represent low distortion. Our analysis is radically different from the analysis carried out in conjunction with task dependent approaches, as we will see in Sect. 4, since we propose a method that evaluates the embedding in an absolute way, instead of evaluating it with respect to how it performs in a given task. We want to highlight the complementarity between our method and the existing ones: our analysis could be performed before the execution of a given

task to select the best (where best means the *least-distorted*) embedding or after task execution as a double check.

Definition 2 (*Distortion*) Given an ontology O , an embedding f and a distance function on O , d_O , the *Average Distortion* is defined as:

$$D(f) = \rho \frac{1}{\binom{n}{2}} \left(\sum_{u,v \in O, u \neq v} \frac{|d(f(u), f(v)) - d_O(u, v)|}{d_O(u, v)} \right)$$

where ρ is a *scaling factor*, defined as: $\rho = \frac{\max d_O}{\max d}$

The scaling factor ρ (also referred to as constant of proportionality) is the normalization factor that makes the distances actually comparable. It represents the ratio between the maximum distances in the two spaces. This metric involves all the pairs of elements that belong to the ontology. In Definition 2, u and v are labels that represent objects both in the ontology and in the embedding. The best case (i.e., the minimum distortion) is $D(f) = 0$, which means that the path lengths are maintained. Even if the embedding distance is very small because the points are placed on top of each other, it could be bigger than 1, but not far from it.

This metric provides a *global* overview of the quality of the embedding, telling us both if the embedding is “good” and how far it is from the perfect representation. The lower bound is 0, which means a zero-distortion representation, while the upper bound is not computable a priori. Generally, distortion values close to 0 are associated with good representations. To provide a complete overview and set an upper bound, we define a complementary metric for evaluating the worst-case distortion.

Definition 3 (*Worst-Case Distortion*) Given an ontology O and an embedding f , the *Worst-Case Distortion* (DWC) is defined as:

$$DWC(f) = \frac{\max_{u,v \in O: u \neq v} d(f(u), f(v)) / d_O(u, v)}{\min_{u,v \in O: u \neq v} d(f(u), f(v)) / d_O(u, v)} \quad (5)$$

That is, the worst-case distortion is the ratio between the maximal expansion and the minimal contraction of distances. The best worst-case distortion is $DWC(f) = 1$, which means that the maximal stretch and the minimal contraction of distances are equal.

4 Case Study: A Knowledge-Driven Geolocation Algorithm

In this section, we present a case study for the application of our embedding-based approach which relates to a knowledge-driven geolocation algorithm (Di Rocco et al. 2018). This algorithm, named Sherloc,⁶ geolocates a microblog message at sub-city

⁶A simple word pun between Sherlock Holmes and location.

level using an external knowledge in the form of an ontology O and its embedded representation S . In the following, we first present Sherlock and then the measure that we used to evaluate the results of our algorithm.

4.1 Sherlock

Given (i) a geographical area of interest, (ii) some external knowledge in the form of a set of objects located in the area of interest and their semantic descriptions, i.e., in the form of an ontology O , (iii) a microblog message originating from the area of interest mentioning at least one of these objects (referred to as *localizable* message) the goal is to infer the coordinates of the mentioned location inside the target area.

In Algorithm 1 (Di Rocco et al. 2018), we present how Sherlock works. Sherlock has five steps, highlighted by comments in the pseudo-code. The input is a message m . We first of all extract from message m only terms that appear in the geographic knowledge (i.e., in the ontology O). Through this process, we obtain a cleaned message $T(m)$ called *geo-message*, which is a set only including the terms in m appearing in O . For each term in $T(m)$ we then compute a k -NN query on the semantic space constructed on O . We recall that the relation between S and O is defined in Sects. 2 and 3. The new message is a set of points in S with maximum cardinality equal to a fixed parameter given as input by the user, called δ .

To infer the coordinates of our message, in step 3, we use the inverse function f^{-1} to convert the geographic terms to their spatial coordinates. This set of points is the input of a clustering algorithm. The collected clusters are then ranked according to density. The densest cluster is the cluster that Sherlock identifies as the cluster of the positions of the message. Finally, to geolocate the message, we compute the convex hull of this cluster. The predicted location is that of the centroid of this convex hull. In order to clarify the Sherlock workflow, we show in Fig. 2 the steps of the algorithm on a real example. The bold terms in the message are the geo-terms that Sherlock processes to infer the coordinates.

There are two important steps that involve the geographic knowledge: k -NN identification and Physical points extraction. More precisely, given a microblog message, Sherlock identifies the nearest neighbor terms closest to the message in terms of semantics, i.e., using the geographic knowledge. After that, Sherlock extracts the physical locations of semantically similar terms again using the geographic knowledge. Sherlock is able to infer the location of a message m without any prior training, exploiting only an indexed geographical external knowledge.

This knowledge-driven solution does not need a training phase. However, this implies that we cannot geolocate every message but only the messages that contain toponyms or geo-terms. These messages are referred to as *localizable* messages.

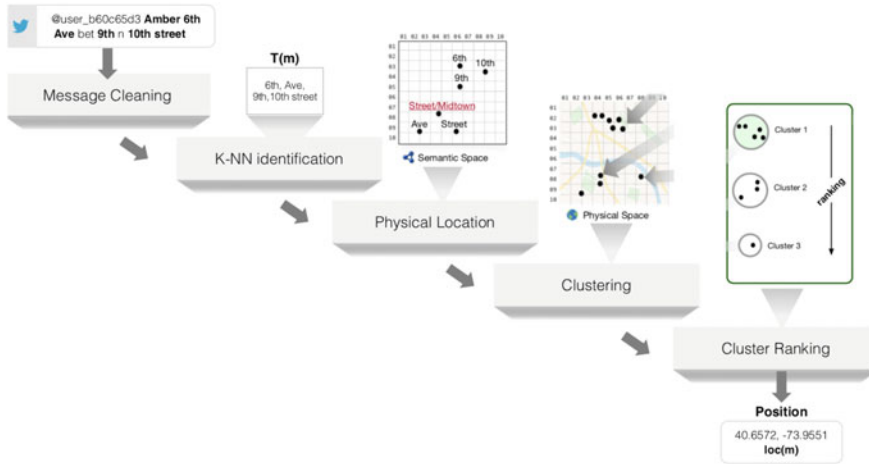


Fig. 2 Sherlock workflow

Algorithm 1: Sherlock

Input: maximum number of similar terms δ , message m , geographic knowledge O , semantic space S

Result: $[lat, lon]$ coordinates of the message m

// Step 1: Cleaning

$T(m) = \text{extract_geo_terms}(m, O)$;

// Step 2: Nearest neighbors identification

$k = \frac{\delta}{\#(T(m))}$;

$NN(m) = \{\}$;

for $t \in T(m)$ **do**

$NN(t) = \text{query}(t, S, k)$;

 // Find NN on S

$NN(m).append>NN(t)$;

end

$T^{NN}(m) = f^{-1}(S, NN(m))$;

// Retrieve closest terms

// Step 3: Physical points extraction

$P(m) = \text{points}(T^{NN}(m), O)$;

// Step 4: Clustering

$\text{Clusters}(m) = \text{clustering}(P(m))$;

// Step 5: Ranking and inferred coordinates

$dC(m) = \max(\text{rf}(\text{Clusters}(m)))$;

$[lat, lon] = \text{centroid}(\text{ConvexHull}(dC(m)))$;

return $[lat, lon]$

4.2 Evaluation Metrics

Given a collection of microblog messages (a Twitter dataset in our specific case) T_w , to evaluate the results of the algorithm we consider T_w as the ground-truth. For the analysis, we choose two commonly adopted distance-based evaluation metrics:

Mean Distance Error (MDE) and Accuracy Distance Error (ADE). Both metrics are defined in terms of the Distance Error $DE(m)$, computed for each message m and defined as the Euclidean distance d between the actual location of m (as in the ground truth), $\text{loc}_r(m)$, and the inferred location, $\text{loc}(m)$: $DE(m) = d(\text{loc}(m), \text{loc}_r(m))$

We choose Euclidean distance for simplicity of computation since the target geographic regions are small and we can approximate the geodesic distance with the Euclidean distance.

MDE is then defined as the average distance error for each message m in the T_w dataset:

$$MDE = \frac{1}{|T_w|} \sum_{m \in T_w} DE(m)$$

5 Experimental Results

In this section, we present the experimental results. We start the discussion introducing the used datasets. Then, we analyze the problem from a task-independent point of view. We conclude the section discussing the results achieved with reference to our specific task, i.e., using the Sherlock geolocation algorithm.

5.1 Datasets

For the experiments, we select information from GeoNames that we use both for the task-independent evaluation and task-dependent evaluation. Moreover, as shown in Sect. 4, we need a Twitter dataset to complete the task-dependent evaluation. Our evaluation is proposed on the target area of Greater London, UK, since the task-dependent evaluation needs to work on a specific target area.

We downloaded the database dump and construct the url for the features using the pattern <http://sws.geonames.org/geonameId/>. A daily GeoNames database dump can be downloaded in the form of a large worldwide text file (`allCountries.zip`). The files are available from the GeoNames server. Since we do not need data of the entire world, we downloaded from the server Great Britain information and then we filtered only the data in Greater London, UK.

With regard to microblog messages, we use a Twitter dataset available online. FollowTheHashtag dataset⁷ contains geotagged tweets retrieved over 167h that correspond to 7 days from 14/04/2016 to 21/04/2016, after removing retweets. No further preprocessing was applied to the retrieved messages. The geographic coverage of this dataset is the entire UK.

In Table 1 we show the summary of relevant information for the evaluation: the target area A that we use on London, described as the bounding box on the city, the

⁷www.followthehashtag.com/datasets/170000-uk-geolocated-tweets-free-twitter-dataset/.

Table 1 Summary of the relevant information used for the evaluation

	London
A	$[-0.5104, 51.2868, 0.334, 51.6919]$
Coverage (km ²)	~ 1100
Max distance (km)	~ 47
m	$m \in \text{FollowTheHashtag}$
O	<i>GeoNames</i>
$\#(T(\text{GeoNames}))$	2209
$\#(Tw)$	44152
Localizable Tweets	33%

Table 2 Distortion values using GeoNames. Bold values represent the best result. S is the embedding space described by its dimensionality. $D(S)$ is the distortion value (see Definition 2) and $std(\cdot)$ is the standard deviation value

S	$D(S)$	$std(D(S))$
3D	0.59842	0.31704
4D	0.61898	0.33013
5D	0.61448	0.32724

coverage of A in km², the max distance in area A . Moreover, we provide a summary of the dataset used showing the cardinality of geo-terms in GeoNames and the number of tweets from the dataset. The last line shows the percentage of localizable tweets that are the tweets that Sherlock can geolocate using GeoNames (since they contain geo-terms).

5.2 Task-Independent Evaluation

In the following, we report the Average Distortion values generated by the different embeddings for the GeoNames ontology (London area). We recall that each ontology is mapped in different N -dimensional embeddings (in our case we consider 3, 4 and 5 dimensions).

In Table 2 we can see that the best representation is produced by 3D embedding. The intuition is the following: *hyperbolic embeddings better project the semantic meaning of an ontology in a low-dimensional space.*

While the 3D embedding is the best, the other dimensions are still suitable to approximate the original structure. This means that the embeddings are able to capture the semantic content of the elements in the ontology and to faithfully represent the distances. The standard deviation values are low, which means that all the representations are quite stable around the mean. The general stability of the different

Table 3 Best MDE results obtained with Sherlock

Sherloc run with S	MDE (in km)
3D	7.28
4D	7.26
5D	7.31

dimensionality representations is due to the fact that the ontology that generates them is the same, and it strengthens the idea that hyperbolic embeddings capture the semantic similarity between geo-terms in a low dimensional space.

It is important to notice that an element in the embedding represents all the semantic meanings of a term. In the evaluation of these results it is also important to underline that the tree representation for the ontology proposed in Sect. 3 is critical in order to capture real paths and logical links among terms.

5.3 Task-Dependent Evaluation

In addition to the distortion analysis, we also use Sherlock as a way of evaluating, in a supervised way, the best embedding dimensions. We use the MDE as a metric to compare the results. In Table 3, we list the best result that we achieve using the different embeddings.⁸ If we compare the results achieved by Sherlock with the results obtained with the distortion value, we can notice that there is coherence between Tables 2 and 3. Sherlock results are always in the same range and, indeed, the distortion value is always less than 1.

6 Conclusions and Future Directions

In this work, we presented a novel approach to represent geographical ontologies (and, in general, hierarchical structures) through word embeddings. Specifically, we focused on the choice of the embedding space in which to project the objects, in order to maintain the semantic (i.e., hierarchical) correlations among them. The spaces that best fit this property are the hyperbolic spaces, specifically the Poincaré disk model. Our evaluation is based on the idea of projecting elements in a space in which also the distances are adequately maintained, introducing the least possible *distortion*. We propose an evaluation that quantifies the distortion (noise) introduced in the representation which maps the geographic ontology onto the embedding. The proposed evaluation schema tackles the numerical problem of comparing objects from different spaces by comparing their reciprocal impact, and introducing a double

⁸Notice that, we run Sherlock with different parameters.

normalization factor. We also propose an analysis of the best data structures, by converting the ontology from a graph to a tree. This conversion emphasizes the property of the Poincaré disk model, because the nature of such space best reflects the tree structure. Since GeoNames can be embedded in another way (Kejriwal and Szekely 2017) we plan to evaluate the effectiveness of this embedding in our work. The quality of the embedding is evaluated in a task-independent way as well as using a real application scenario, i.e., a geolocation algorithm.

The results obtained with both evaluations are very good and reinforce our hypothesis that using embeddings for representing geographical ontologies is a viable approach. Moreover, we can observe that there is coherence between the results of the task-independent and the task-dependent evaluation.

Since our results demonstrate that external geographic knowledge embedded in a metric space provides a good solution to easily find distances between points, we plan to analyze also different geographic data sources in both the task-independent and task-dependent evaluation. Moreover, we want to test different embeddings algorithms and try to assess their respective effectiveness in representing geographic ontologies.

References

- Bakarov A (2018) A survey of word embeddings evaluation methods. [arXiv:1801.09536](https://arxiv.org/abs/1801.09536)
- Baroni M, Dinu G, Kruszewski G (2014) Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp 238–247
- Bengio Y et al (2003) A neural probabilistic language model. *J Mach Learn Res* 3:1137–1155
- Bojanowski P et al (2017) Enriching word vectors with subword information. *Trans Assoc Comput Linguist* 5:135–146
- Di Rocco L (2018) The role of geographic knowledge in sub-city level geolocation algorithms. PhD thesis. Dibris, Università degli Studi di Genova
- Dredze M, Osborne M, Kambadur P (2016) Geolocation for twitter: timing matters. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp 1064–1069
- Faruqui M et al (2016) Problems with evaluation of word embeddings using word similarity tasks. In: Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP, pp 30–35
- Griffiths TL, Tenenbaum JB, Steyvers M (2007) Topics in semantic representation. *Psychol Rev* 114(2):211–244
- Gruber TR (1993) A translation approach to portable ontology specifications. *Knowl Acquis* 5(2):199–220
- Harris ZS (1954) Distributional structure. *WORD* 10(23):146–162
- Jia Y et al (2014) Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp 675–678
- Kejriwal M, Szekely P (2017) Neural embeddings for populated geonames locations. *The Semantic Web - ISWC 2017*, pp 139–146
- Krioukov D et al (2010) Hyperbolic geometry of complex networks. *Phys Rev E* 82(3):036106-01–036106-18

- Mikolov T et al (2013) Efficient estimation of word representations in vector space. In: ICLR Workshop
- Nayak N, Angeli G, Manning CD (2016) Evaluating word embeddings using a representative suite of practical tasks. In: Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP, pp 19–23
- Nickel M, Kiela D (2017) Poincaré embeddings for learning hierarchical representations. *Adv Neural Inf Process Syst* 30:6338–6347
- Nickel M, Kiela D (2018) Learning continuous hierarchies in the Lorentz model of hyperbolic geometry. In: Proceedings of the 35th International Conference on Machine Learning, pp 3776–3785
- Pennington J, Socher R, Manning CD (2014) Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pp 1532–1543
- Sala F et al (2018) Representation tradeoffs for hyperbolic embeddings. In: Proceedings of the 35th International Conference on Machine Learning, pp 4460–4469
- Sayeed AB, Greenberg C, Demberg V (2016) Thematic fit evaluation: an aspect of selectional preferences. In: Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP, pp 99–105
- Schnabel T et al (2015) Evaluation methods for unsupervised word embeddings. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp 298–307
- Stadler C et al (2012) LinkedGeoData: a core for a web of spatial open data. *Semant Web* 3(4):333–354
- Wang Q et al (2017) Knowledge graph embedding: a survey of approaches and applications. *IEEE Trans Knowl Data Eng* 29(12):2724–2743
- Zheng X, Han J, Sun A (2018) A survey of location prediction on Twitter. *IEEE Trans Knowl Data Eng* 30(11):1652–1671

Web-Based Visualization of Big Geospatial Vector Data



Florian Zouhar and Julia Senner

Abstract Today, big data is one of the most challenging topics in computer science. To give customers, developers or domain experts an overview of their data, it needs to be visualized. In case data contains geospatial information, it becomes more difficult, because most users have a well-trained experience how to explore geographic information. A common map interface allows users zooming and panning to explore the whole dataset. This paper focuses on an approach to visualize huge sets of geospatial data in modern web browsers along with maintaining a dynamic tile tree. The contribution of this work is, to make it possible to render over one million polygons integrated in a modern web application by using 2D Vector Tiles. A major challenge is the map interface providing interaction features such as data-driven filtering and styling of vector data for intuitive data exploration. A web application requests, handles and renders the vector tiles. Such an application has to keep its responsiveness for a better user experience. Our approach to build and maintain the tile tree database provides an interface to import new data and more valuable a flexible way to request Vector Tiles. This is important to face the issues regarding memory allocation in modern web applications.

Keywords Bigdata · Visualization · Vector-tiling · Geospatial data · Web

F. Zouhar (✉) · J. Senner
Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany
e-mail: florian.zouhar@igd.fraunhofer.de

Technische Universität Darmstadt, Darmstadt, Germany

J. Senner
e-mail: julia.senner@igd.fraunhofer.de

© Springer Nature Switzerland AG 2020
P. Kyriakidis et al. (eds.), *Geospatial Technologies for Local and Regional Development*, Lecture Notes in Geoinformation and Cartography,
https://doi.org/10.1007/978-3-030-14745-7_4

1 Introduction

Utilizing modern technologies, Computer Scientists are able to store a massive amount of data of all kinds. Data is used to create statistics and analyze behaviors of machines, humans, the nature and much more. Algorithms from data aggregation to machine learning are applied to gain more information and to interpret this amount of data.

Adding geospatial references to data makes it more valuable. Users, domain experts, developers and scientist have a better intuition of data, when handling it in a geographic context on earth. It is an essential topic to visualize data not only as statistic diagrams or complex interactive graphs. Geospatial data should be visualized on a map to relate to the geographic context. Users can explore the data by panning and zooming like they are used to by applications such as Google Maps¹ or OpenStreetMap.² Users are well-trained on using these interfaces to explore countries, streets, cities, buildings or satellite images.

In order to build such a map interface, modern application make use of web technology frameworks like OpenLayers.³ They implement two different techniques. The most common approach used for street maps is to slice the world into tiles, render raster images and transmit them to the browser. The other option is loading vector data, and render it using application or user defined styling dynamically. This way, more interactive applications are possible, such as directly filtering and modifying features. The main challenge here is to visualize a massive amount of geometries without losing render performance and interface responsiveness, hence user experience.

State of the art solutions mostly deal with static data which makes a visualization much easier as streets or buildings are changing rarely. They pre-render the whole world, which is a very time-consuming task and this method cannot be applied on fast changing data.

The existing standard solution to load vector data into a browser is a Web Feature Service (WFS)⁴ implementation. Users explicitly define which part of the world and which type of data they want to explore. Having spatially dense data, distributed over a large area, cannot be explored interactively using this technology, because loading the initial map would query data for the whole world. One advanced example is visualizing information about growing crops in millions of parcels spread over the country. The information about the growing plants, trees and grain types are updated periodically, which makes the data dynamic. The main challenge is to provide a map interface to explore this amount of dense data using a vector based approach.

The contribution of this work is an approach which yields the benefits of both solutions mentioned before. It is important to provide interaction features such as exploring temporal datasets or apply data-driven styling without reloading data from

¹ Google Maps—<https://maps.google.com>.

² OSM—<https://www.openstreetmap.org/>.

³ OpenLayers—<http://openlayers.de>.

⁴ OGC WFS—<http://www.opengeospatial.org/standards/wfs>.

servers. In particular, our approach makes it possible to render over one million polygons integrated in a modern web application along with user defined styling.

This work focuses on vector based tiling approaches and suitable JavaScript frameworks for layer based map applications. It is not about introducing a geographic information system for the web or implementing a cloud processing solution.

There are three essential steps this solution is providing. At first, data has to be stored in a way to have efficient geospatial access. This is achieved by using a geospatial index along with a fast and scalable distributed filesystem.

Secondly, the data has to be optimized for visualization purposes. This is done by implementing a tiling algorithm and transforming the data to a specialized format. This format has the benefits of faster visualization and network transmission while efficiently saving storage space.

Finally, data has to be transmitted to a web application running in modern web browsers. The geometries then are rendered using a WebGL map application framework. Furthermore, it is possible to add interaction concepts such as filters and user defined styling. Additionally, the idea is to manipulate, delete and add geometries directly.

The geospatial data to be processed and visualized comes as files in a textual format. The approaches are made clear using a real world use case. Agronomists and insurance companies have to monitor the growing process of crops to make important decisions. To perform well, data is gained for parcels in the whole country. These huge datasets contain geometries describing agriculture parcel boundaries and attributes, describing the type and the actual health of a parcel. This data needs to be visualized in a way also non geographic experts can explore it. In addition to that, interaction should be implemented, such as time-range based filters or data-driven colorization of the parcels.

2 Related Work

Research in the context of tiling mostly focus on raster image based approaches, rather than making big vector data accessible to web applications. This section gives an overview on existing methods in storing, processing and visualizing vector data in a geospatial context. Vitolo et al. (2015) has written a survey on how bigdata is handled using web technologies. In evolution of computer science data collections expand very fast. New technologies are needed in order to handle such a huge amount of data. He describes this has to be possible in context of analysis, workflows and interaction within this datasets. The paper discusses the current web technologies used to process simple dataset, which can be used very straight forward. In addition to that current web technologies implementing standards maintained by the Open Geospatial Consortium,⁵ are lacking of flexibility and scalability.

⁵ Open Geospatial Consortium—<https://www.opengeospatial.org/>.

While Vitolo was focused on bigdata and web technologies (Yue and Jiang 2014) discusses big data in the context of geospatial information systems. In his opinion it is important to support the geospatial domain as especially GIS software should be able to handle huge datasets containing geospatial data. Concepts from Horak et al. (2010) in *Web Tools for Geospatial Data Management* focuses on providing remote interfaces based on XML and defined also by the OGC. He discusses the data flow from databases or files to web based applications. But it is not really focused on bigdata and the aspect of exploring the data using a visualization.

The format used to store and transmit data is important. Standards are defined using plain text formats such as JSON and XML. These are not efficient to store or transmit without compression and encoding. Data format research is mostly done to do faster transmission of spatial data over the network. One approach by Yang and Li (2009) is data compression by clustering data. The ideas are to compress data, but not for a visualization purpose. Tiling is also discussed in the using XML or JSON based data, which is transmitted, and merged in order to render vector data using SVG (Antoniou et al. 2009). In the context of tiling, caching is a widely mentioned topic (Liu et al. 2007). Also, Blower (2010) is discussing caching of raster tiles and GIS in the Cloud. Ingensand et al. (n.d.) and van den Brink et al. (2017) as well as many others focusing on raster image tiling. Creating raster tiles on the fly server side is a huge topic as well. Olasz et al. (2016) has written a survey on the possibilities on using server-side rendering. In particular he uses GeoTrellis⁶ and several different storage solutions to evaluate in which context these technologies can be used. Visualization in the manner of bigdata here is a small topic. Just creating a tiling tree with raster images is mentioned. As described for now the state of the art at the moment is to render tiles on servers and accept the lack of interaction possibilities. This concept of tiling and storing raster images has to be adapted to upcoming requirements utilizing the features of vector tiles.

Ideas of tiling to provide a map user interface are really focusing on tiling but not on trying to compress, cache tiles in a vector format without increasing memory usage at any point in the workflow.

The only stable data format for 2D vector tiling currently is Mapbox Vector Tiles. Eriksson and Rydkvist (2015) has done review on the map render framework implemented by Mapbox called Mapbox GL JS. It focuses on a huge number of features and the performance of the render engine implemented by Mapbox. The tile generation is done by using a command line tool after exporting a GeoJSON file from PostGIS. It relies on building a tile set once a time without providing capabilities like adding data. The tiling itself is the focus, rather than optimizing compression, encoding and caching of vector tiles.

Vector Tiles introduced by Mapbox are using Google Protobuf. This is an efficient way to serialize and transmit structured data over a network. Feng and Li (2013) describes usage in the context of online games. Mapbox provides a detailed specification of Vector Tiles using Protobuf encoding. One should go further and evaluate if there are better solutions in order to be more flexible in working with vector tiles

⁶ GeoTrellis—<https://geotrellis.io>.

and dynamic data. This is important to build applications processing and creating Vector Tiles dynamically.

Scalable Vector Graphics (SVG) were used as part of the development of spatial data visualizations in browsers. Visualization ideas using SVG (Langfeld et al. 2008) are straight forward because the rendering task is handled by the browser's engine. But SVG is not able to handle huge amounts of geometries, as everything is loaded into the browsers Document Object Model (DOM). Rendering directly onto a canvas is way more efficient and has replaced the SVG technology for rendering purposes mostly.

To sum up, there is a lot work done focusing on data standards and storing data using state of the art technologies. Server components focusing mostly on raster data. It is important to evaluate the usage of vector data in manner of geospatial visualizations and big datasets.

The next paragraphs give a brief overview on the mostly used standard database and publishing service for geospatial data. This is a relevant topic because working and optimizing data relies on a data storing solution.

2.1 *PostGIS*

Relational databases are the most spread and traditional way to store structured data besides a plain file system. SQL is used to maintain and access data in a PostgreSQL⁷ Database. It supports the definition of custom functions and the implementation of extensions. Extensions have to be written especially for PostgreSQL. These are written using a combination of SQL statements, C code snippets and configuration files. Maintaining and extending an application based on PostgreSQL is hard to manage because of the complexity and the limitation of the architecture.

PostGIS⁸ is a feature rich and complex extension for PostgreSQL Databases to store and index geospatial data. PostGIS stores Geometries in a spatial index in Well Known Text⁹ format along with the given properties. The evaluation section will show how PostGIS performs on importing huge GeoJSON files with on the fly conversion to WKT.

There are many functions in SQL syntax to query, aggregate and convert data which is stored in such a database. Mapbox Vector Tile format is a supported export format as well. Every tile is created on demand, which lacks of performance on thousands of polygons per tile. Three main PostGIS functions have to be nested in order to create a tile. Creating such a custom SQL functions to get tiles directly from the database is a very uncommon way. Additional functionality has to be

⁷ PostgreSQL—<https://www.postgresql.org/>.

⁸ PostGIS—<https://postgis.net>.

⁹ WKT—<https://www.opengeospatial.org/standards/wkt-crs>.

implemented in form of a middleware providing an interface handling the SQL queries. The next paragraph describes the functions of GeoServer which is implemented for this purpose.

2.2 *GeoServer*

Based on a PostGIS database, GeoServer¹⁰ can publish the data by providing several standardized interfaces. The OGC standards WFS, WMTS and WMS are supported. The Web Map Tiling Service (WMTS) also serves tiles in Vector Tile format. It uses its own implementation to serve Vector Tiles and creating a cache. What it simply does when requesting a tile is creating a query including the bounding box for the requested tile and then using a Vector Tile implementation by Electronic Chart Centre¹¹ to clip and encode it. These tiles are cached directly in a file system. Surprisingly, creating vector tiles takes longer than actual raster images. The vector tile creating is not optimized for years like the creation of raster tiles. Depending on the amount of data the creation of a tile can take several minutes.

Using GeoServer, one has no control which data is included in the vector tiles. Every attribute stored in the particular PostGIS table is attached to the geometries. In manner of caching one has only two control mechanisms to invalidate tiles. The first is to define a fixed timeout, the other option is to manually trigger a rebuild on the whole world or a particular region.

3 Our Approach

This section describes our implementation approach in detail. The idea consists of three essential components.

- Storage 3.1
- Tiling Approach
 - Geometry Processing 3.2.1
 - Vector Tile Encoding & Storage 3.2.2
- Visualization 3.3

Data is maintained in the storage component. The main contribution in this work is the implementation of the processing done in order to build, extend vector tiles and provide a API to request these tiles. Lastly the focus is to visualize the geometries using a reliable map application framework, while adding interaction features to explore the data.

¹⁰ GeoServer—<http://geoserver.org/>.

¹¹ ECC Vector Tile Encoder—<https://github.com/ElectronicChartCentre/java-vector-tile>.

All used technologies and libraries are written in JVM languages, as well as our implementation. The focus is on facing the research objectives, rather than dealing with technical details. For this reason, the high-level language Kotlin is used for implementation. Kotlin is a relatively new language founded by JetBrains and glues benefits from Java and Scala together. Kotlin supports object oriented and functional programming techniques.

The next paragraphs describe the components in detail.

3.1 Storage

In order to import, query and aggregate geospatial data, this approach uses GeoRocket¹² as storage technology. GeoRocket is a high-performance data store for geospatial files. It uses MongoDB¹³ to persist data and Elasticsearch¹⁴ to build a spatial index for data query and aggregation tasks. GeoJSON can be imported directly without conversion.

3.2 Tiling Approach

The following two objectives in the manner of vector tiling are faced here. Assume new data is imported into the data store. The tiling component has to fetch this data and merge it into the existing tile tree database. Secondly, requests on tiles have to be processed in a reasonable amount of time. These requests contain the actual coordinate of the tile and a query, which information explicitly should be integrated into the tiles. This is also the core contribution of our approach.

The tiling implementation is a server component itself and provides a REST interface. It can be configured using different file storage backend technologies for persisting the tiles. The configuration includes a range of zoom levels in which the tiles are created, which is 2 to 15 by default. These are sufficient enough for most users map interface experience. For a more detailed view tiles on higher zoom levels can be build.

3.2.1 Geometry Processing

Assume a GeoJSON file read to working memory ready to be processed. This section describes how a collection of geospatial features is processed to create the actual tile afterwards.

¹² GeoRocket—<https://georocket.io>.

¹³ MongoDB—<https://www.mongodb.com/>.

¹⁴ Elasticsearch—<https://www.elastic.co/products/elasticsearch>.

For further processing, geometry coordinates are Mercator projected to values from zero to one [0, 1]. At first the bounding box of all geometries is calculated. This computation is important as it saves a lot of processing time in the end. In order to maximize the ability of parallel computation, every zoom level can be processed independently in our algorithm. On every zoom level, the bounding box is used to determine which tiles have to be created. The iteration starts creating the top-left tile covered by the bounding box, which can be calculated for a certain zoom level z and the top-left bounding box coordinates x, y (see Eq. 1).

$$(2^z * x, 2^z * y) \quad (1)$$

For each covered tile a feature collection is created. All geometries are clipped at the tile edges. Clipping defines the process of cutting geometries off at some boundaries to save disk space and improve render performance. To reduce polygon complexity, the Douglas and Peucker (1973) algorithm for line simplification can be applied. It is a lightweight process to remove spikes and small corners using a precision parameter. On lower zoom levels these are too small to be recognized by a user. The last step before building the actual tile transforms the generic world coordinates into the tiles extend, which is 0 to 4096 by default. Equation 2 shows how the transformation is done.

$$(\lfloor extend * (xVal * (1 \ll z) - x) \rfloor, \lfloor extend * (yVal * (1 \ll z) - y) \rfloor) \quad (2)$$

Listing 1 illustrates the algorithm. The here stated algorithm should not provide a highly optimized solution, rather than showing that even our simple approach performs very good in creating tiles. See the evaluation Sect. 4 for details. Our implementation provides linear scalability as every zoom level processing runs in its own thread.

3.2.2 Vector Tile Encoding and Storage

The second and most important sub-component of the tiling component is the implementation of storing and serving vector tiles. Vector tiles are encoded using Google Protobuf and afterwards compressed using GZip. A storage interface is implemented supporting several backend technologies, such as MongoDB, SQLite and simply a directory structure using the native file system. The binary encoded tiles can either be stored using a triple primary key schema (z, x, y) or a unique hashed value when stored into a MongoDB Collection.

The following describes how a complete GeoJSON collection is encoded and stored. The geometries are already clipped at the tile boundaries and coordinates are transformed accordingly to the vector tile specification. In order to encode the data in Protobuf format the properties have to be transformed. The simple key-value map is converted into a tag-set. A tag-set is a data structure consisting of a distinct set containing keys and values. Features are then tagged using the indices in this

Algorithm 1 Tiling Algorithm

```

MIN_ZOOM ← 2
MAX_ZOOM ← 15
EXTENT ← 4096
BUFFER ← 64
CPUS ← 2
BULK ← 500_000
BBOX ← WORLD

function TILE(InputFeatures)
  while FeatureCollection = GETFEATURES(InputFeatures, BULK) do
    BBOX ← CALCBOUNDINGBOX(FeatureCollection)
    for z in MIN_ZOOM to MAX_ZOOM do
      LAUNCH.TRAVERSEZOOM(FeatureCollection, z)
    end for
  end while
end function

function TRAVERSEZOOM(InputFeatures, ZoomLevel)
  Q ← GETCOVEREDTILES(ZoomLevel, BBOX)
  while tile ← Q.NEXT do
    clippedFeatures ← CLIP(InputFeatures, Tile)
    CREATEANDSTORE(clippedFeatures)
  end while
end function

```

set. There exist multiple open source libraries like the ECC Vector Tile Encode¹⁵ to parse in memory data to the Protobuf schema. Polygons are transformed to polylines, which also saves disk space and render time. Polylines only store the actual vector rather than the absolute coordinates. This official Protobuf¹⁶ encoder writes the actual binary, which is compressed using GZip and stored into the database.

Inserting a feature collection into an existing tile tree needs more explanation. One needs to check whether a tile exists to prevent overriding it with the new one. When the old tile exists, it is loaded into memory. Depending on the number of features and attached properties this can consume huge amounts of memory. Then the geometry lists are concatenated and both tag-sets have to be merged. To merge two tag-sets a iteration over the tags and indices has to be done. This is the reason why the whole tile is extracted into memory. Once this is done, the Protobuf encoder does its job encoding the tile and the old one can be overwritten. This process enables us to insert new data into a vector tiles tree, which is an important and simple feature. The main benefit here is, state of the are map interface frameworks for web browsers are able to read this format directly without any modification.

Our Tiler provides a REST interface to import a file, clear the tile database and request a single tile in multiple formats. The following endpoints are available.

¹⁵ ECC Vector Tile Encoder—<https://github.com/ElectronicChartCentre/java-vector-tile>.

¹⁶ Protobuf GitHub—<https://github.com/protocolbuffers/protobuf/tree/master/java>.

- *POST* /—Import GeoJSON
- *DELETE* /—Clear Database
- *GET* / : z / : x : / : y { . pbf , . mvt , . geojson }—Request a Tile

Using this tile endpoint could serve a complete tile including all geometries and attributes. Depending on the number of features per tile, these can reach over several megabytes in size even when encoded and compressed. Another main reason for space consuming tiles are geometry attributes. To fetch only necessary data, only a unique identifier per feature and the actual geometries are included into the tile. The identifier is needed to refer to actual data in the storage. The following parameters for tile requests can be used for a more flexible tile request.

- *filters*—a key-value map to include only features matching these attribute values
- *bounding_boxes*—an array of bounding boxes in WGS 84. Only tiles and geometries matching or intersecting these boxes are served.
- *fields*—an array of attribute keys which should be included into the features in any case

Once a tile is requested it will be loaded from the database into memory parsed into a object model for the ease of processing. Then features matching the criteria in *filters* and *bounding_boxes* and only requested attribute fields are collected. The encoder then does its job and streams the compressed resulting tile to the client.

This tile processing is very important to save transmission time and memory space in browsers. Figure 1 illustrates the request process. The next section describes the visualization itself and shows why that costly process is important.

3.3 Visualization

Finally, the created tiles can be integrated into a map interface implemented for modern web browsers. Mostly JavaScript frameworks are used to provide such an

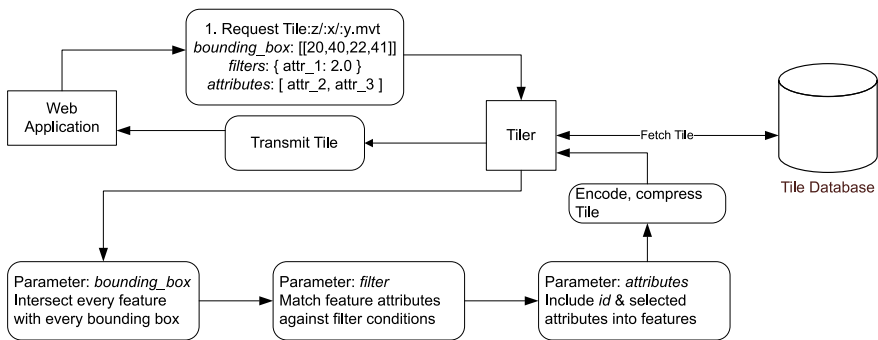


Fig. 1 Illustration of a dynamic tile request

interface. The most common and stable frameworks are OpenLayers¹⁷ and Mapbox GL JS.¹⁸ The young vector tile implementation in OpenLayers has many issues, most critical a memory leak, no data-driven styling and no WebGL support for vector tiles.¹⁹ Therefore, Mapbox GL JS was used for the evaluation. As base layer we added an OpenStreetMap default raster layer.

In our use case the vector tiles contain one layer only consisting of polygons. Our implemented tile endpoint in $z/x/y$ format is simply added as additional layer in our map interface. Now the tiles are rendered on top of the OpenStreetMap base layer. At this point users already can explore their data using the map interface. We managed to render about two million polygons, which is enough in most use cases. Humans can not differentiate such a number of polygons rendered on top of a street map, but it gives a good overview. For a more detailed few into the data we provide interaction features. Users can select single geometries to inspect their attributes. These are loaded using the previous mentioned identifier from the storage solution. Furthermore, this user interfaces is extended with filtering and colorization capabilities. Data can be filtered by different conditions, such as time ranges, numeric values or discrete keywords. Data-driven styling can be applied to the geometry fillings in correlation with numeric attributes. To do so the range between the minimum and the maximum of the particular attribute is interpolated onto a user defined color gradient. The user interface also allows to draw rectangles added as a bounding box filter condition. This can minimize the amount of features per tile and the amount attributes per features. As mentioned before the system is very flexible. It is important that any user interface can be built on top of our endpoint. Figure 2 shows a visualization of the data set used for evaluation with applied data-driven colorization.

The next section shows the core benefits of this approach and the limitations.

4 Evaluation

Every component from our approach needs an evaluation. The following measurements and the core contribution features are evaluated.

- Storage Import Times 4.1
- Tiling Process Measurement 4.2
- Visualization and Flexible Tile Request 4.3

¹⁷ OpenLayers—<http://openlayers.de>.

¹⁸ Mapbox GL JS—<https://mapbox.com>.

¹⁹ New versions may be released since, we used version 4.0.

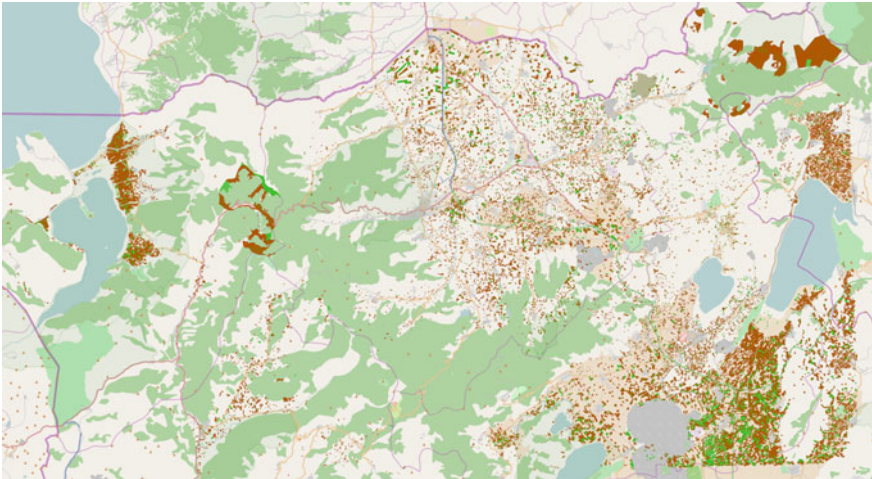


Fig. 2 Visualization of about 500,000 parcels

Table 1 Importing GeoJSON to storage

#Polygons	Import PostGIS	Import GeoRocket
100k	1 min	32 s
250k	3 min 27 s	1 min 21 s
500k	4 min 38 s	2 min 38 s
1 mi	9 min	5 min 39 s

4.1 Importing to the Storage

For our purposes the GeoJSON files were imported into PostGIS and GeoRocket, each containing a fixed number of features. Each feature has a unique identifier and one medium complex Polygon. Also it contains ten attributes, which includes numeric, date and string values. Table 1 compares the processing times of both technologies. These databases are used to maintain the data, query original data portions and also to perform data analytic tasks like aggregations. It is important to mention the storage in this context, because it is essential for a proper data management. Importing times scale linear as both databases import and index data feature by feature. These timings are important because it shows how long it takes on a database until the actual visualization building process is started.

After importing one file GeoServer has to rebuild the covered area completely. It is important to understand that we use GeoRocket only as storage for maintenance purposes and additionally to load feature information explicitly or perform data analytic tasks like aggregations. The tiling process starts directly upon import. Therefore, a user will see the first zoom levels much faster than waiting for the PostGIS database.

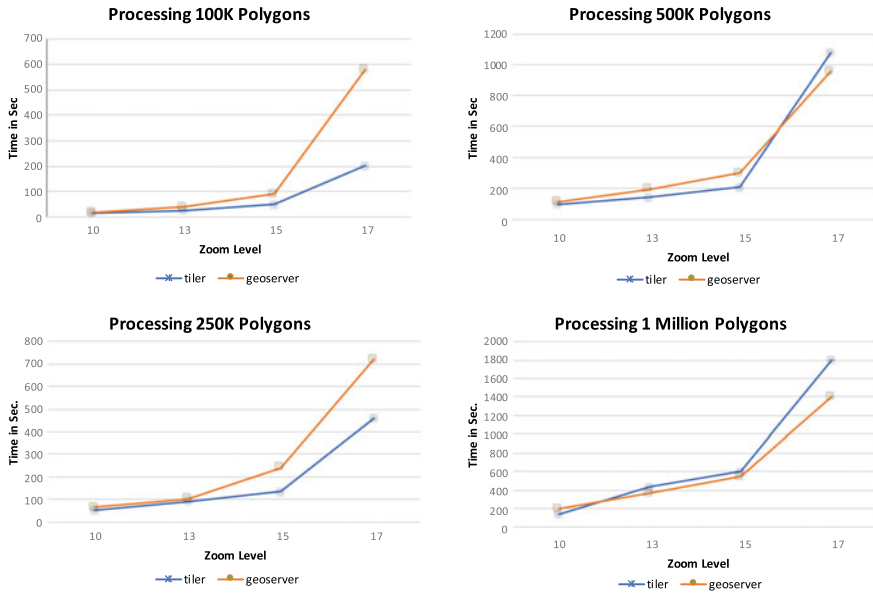


Fig. 3 Processing of different sized files

4.2 Tiling Process Measurement

In order to compare GeoServer and our tiling approach, performance measurements are done on GeoJSON files as described in the importing process Sect. 4.1. For each file different zoom level ranges were configured in order to show the time needed to process higher zoom levels. Creating zoom levels above 15 turns out to be much slower. The reason for this is, on higher zoom levels a massive number of tiles, respectively files, have to be stored to disk. Depending on the region covered, it can be billions on level 17. This evaluation uses two thread threads to create one zoom level each. Also, a bulk size of 500,000 features is processed per thread, which prevent the system to allocate to much memory. Once a feature collection is ready to build an actual vector tile, it is passed to the encoder writing the actual binary to the tile database. Figure 3 compares tile set creation times using our approach and GeoServer. Tile creation or merging is faster using our approach for smaller datasets. This is a benefit when smaller data portions are imported continuously. In addition to that our approach is flexible. By adjusting the bulk insertion size, one can decide to buffer more or less, depending on the data import flow.

4.3 *Visualization and Flexible Tile Request*

The important part is to transmit tiles over a network and provide a map interface visualizing the geometries. To be mentioned first, our interface can handle up to two million polygons using vector tiles. This is sufficient number as humans will not be able to differentiate these many objects in a map interface. This only gives an overview which parts on earth are covered with data. The biggest tiles in our dataset, including multiple thousand features and the before mentioned ten attributes, reaches about 15 megabytes in size after being compressed.

In this work the main contribution is the not only the processing of tiles, but also serving these upon request. A user interface may provide a data overview while a user afterwards decides what data to explore further. In our case the user applies a filter condition on one attribute and also enables a custom styling depending on two attributes. This information is attached to a tile request. Figure 1 illustrated this request and further processing. The tiling server then decodes the requested tile, applies the filter and includes only the identifier along with two attributes needed for the styling to every feature and afterwards encodes the tile before serving it to the client. The benefit is, this system is very flexible. It lets the user or interface developer decide which data should be included into a tile, besides styling geometries as desired. Requesting all feature information can be done at any time using the reference to the storage. Keeping features clear from too much attributes speeds up transmission time and keeps memory allocation in browsers low. Rendering a huge number of geometries is no problem in our interface, but holding many keyword attributes for every feature, such as the crop types, in memory is very costly. These attributes can be loaded on demand if needed from the spatial index database. This also reveals the limitation when loading the complete tiles into a modern browser. All interaction features like filtering, styling and geometry inspecting can be done without reloading data, but this also allocates too much memory. This shows the contribution in context of BigData and modern web browsers. This can easily reach more than two gigabytes, which is also a hard limit for most browsers to allocate in one tab. Secondly, processing huge tiles to apply filter conditions is a costly task, which also needs a deeper insight.

5 Conclusion

To sum up, our approach shows processing data continuously to create a vector tile database can be done fast without recreating whole parts of the tile tree every time new data is imported to the storage. Also, we showed there are possibilities to render huge amounts of geometries. In order to provide full interaction some more processing has to be done, but this also shows the flexibility of the system.

The processing times in GeoServer and our Tiling Component are very similar for huge amounts of features at once. But the real benefit relies in the workflow how

tiles are handled. First of all, the GeoServer starts to build tiles after data is imported into the PostGIS database (see Table 1). In our approach a user will see the first zoom level much earlier, because importing can be done straight away. We figured out zoom levels 2 to 15 are sufficient for the most use cases. Zoom level 2 gives a good overview, while 15 has enough details. The second and more important limitation using the GeoServer is, one has to rebuild tiles completely for the region covered by the newly imported data. A request to PostGIS querying all features for a tiles specific region is performed in order to create the tile. Our approach allows to extend not only the tile tree, but also the tiles itself. In our use case data is imported every second week. Recreating a whole tile would lead to a time-consuming task. In our approach a tile is updated. Once a feature collection is ready, the lastly saved tile is loaded, and merged with the new data. Finally, its saved to disk again. As encoding is the fastest part of the processing, creating fresh tiles and merging data into existing ones does not make a huge difference. The limiting factor here is memory space needed for the uncompressed tiles to merge or recreate them. Our implementation is not much optimized yet, especially in the manner of memory usage. It shows it is possible to maintain a vector tile database in a reasonable amount of time.

Another huge benefit to use only vector data is, the limitation of the image resolution is no longer a problem. Vector data can be transmitted faster and the rendering does not lack of quality.

All performance measurements are done on a MacBook Pro 13-inch (2017, 16 GB memory, Core i5 3.1 Ghz processor). This hardware should represent a average mid-level hardware. It has a high resolution display, which also makes the benefit of the rendering quality using vector data visible.

6 Future Work

The evaluation section reveals the limitations of working with state of the art vector tile technologies. The idea is to keep track on our approach to speed up the vector tile creation process which is important for both, the initial creation of the tile tree and also serving tiles. Memory allocation during the decoding of a whole tile is huge limitation. This can easily be five gigabytes per tile. Google Protobuf is conceptual designed in a way one has to decode the whole message to read it. Our next step could be to have an internal encoding to just skip parts of our tile while reading it. Features or attributes not requested are skipped then. Furthermore, utilizing this fast processing could give us the option to apply culling. We figured out, especially in time series data, many geometries are the same and therefore overlap. These could also be skipped by default. On the visualization side we want to give a fast and simple overview of the data in a spatial manner. The main idea here is to do more simplification on low levels of detail. We consider to apply clustering and merging of geometries. Additionally, in a more detailed few the idea is to make even direct²⁰

²⁰ DataBio—<https://datbio.eu>.

and persistent geometry manipulation possible. We are also working on technical optimizations in our algorithms²¹ and technologies.

Acknowledgements Research presented here is carried out within the data-driven bioeconomy project Databio. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 732064. It is also part of the Big Data Value Public-Private Partnership. We would like to thank Prof. Dr. Ir. Arjan Kuijper for his valuable comments and input.

References

- Antoniou V, Morley J, Haklay MM (2009) Tiled vectors: a method for vector transmission over the web. In: Carswell JD, Fotheringham AS, McArdle G (eds) *Web and wireless geographical information systems*. Springer, Heidelberg, pp 56–71
- Blower, JD (2010) GIS in the cloud: implementing a web map service on google app engine. In: *Proceedings of the 1st international conference and exhibition on computing for geospatial research and application, COM.Geo '10*, ACM, New York, NY, USA, pp 34:1–34:4
- Douglas D, Peucker T (1973) Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartogr: Int J Geogr Inf Geovisualization* 10:112–122
- Eriksson O, Rydqvist E (2015) An in-depth analysis of dynamically rendered vector-based maps with WebGL using Mapbox GL JS. Master's thesis, Linköping University Linköping University, Software and Systems, Faculty of Science & Engineering
- Feng J, Li J (2013) Google protocol buffers research and application in online game. In: *IEEE conference anthology*, pp 1–4
- Horak P, Charvat K, Vlk M (2010) *Web tools for geospatial data management*. Springer, Boston, pp 793–800
- Ingensand J, Nappez M, Moullet C, Gasser, L, Ertz O, Composto S (n.d.) Implementation of tiled vector services: a case study
- Langfeld D, Kunze R, Vornberger O (2008) SVG web mapping. Four-dimensional visualization of time-and geobased data
- Liu Z, Pierce ME, Fox GC, Devadasan N (2007) Implementing a caching and tiling map server: a web 2.0 case study. In: *2007 International symposium on collaborative technologies and systems*. pp 247–256
- Olasz A, Thai BN, Kristóf D (2016) A new initiative for tiling, stitching and processing geospatial big data in distributed computing environments. *ISPRS Ann Photogramm, Remote Sens Spat Inf Sci* 3:111
- van den Brink L, Barnaghi P, Tandy J, Atemezing G, Atkinson R, Cochrane B, Fathy Y, Castro RG, Haller A, Harth A et al (2017) Best practices for publishing, retrieving, and using spatial data on the web
- Vitolo C, Elkhatib Y, Reusser D, Macleod CJ, Buytaert W (2015) Web technologies for environmental big data. *Environ Model Softw* 63:185–198
- Yang B, Li Q (2009) Efficient compression of vector data map based on a clustering model. *Geo-Spat Inf Sci* 12(1):13–17
- Yue P, Jiang L (2014) BigGIS: how big data can shape next-generation GIS. In: *2014 The third international conference on Agro-Geoinformatics*. pp 1–6

²¹ BDV PPP—<http://www.bdva.eu/PPP>.

Part II
Geoinformation Science and Geospatial
Technologies in Transportation

A Clustering-Based Framework for Understanding Individuals' Travel Mode Choice Behavior



Pengxiang Zhao, Dominik Bucher, Henry Martin and Martin Raubal

Abstract Travel mode choice analysis is a central aspect of understanding human mobility and plays an important role in urban transportation and planning. The emergence of passively recorded movement data with spatio-temporal and semantic information offers opportunities for uncovering individuals' travel mode choice behavior. Considering that many of these choices are highly regular and are performed in similar manners by different groups of people, it is desirable to identify these groups and their characteristic behavior (e.g. for educational or political incentives or to find environmentally-friendly people). Previous research mainly grouped people according to "mobility snapshots", i.e. mobility patterns exhibited at a single point in time. We argue that especially when considering the change of behavior over time, we need to investigate the behavioral dynamic processes resp. the change of travel mode choices over time. We present a framework that can be used to cluster people according to the dynamics of their travel mode choice behavior, based on automatically tracked GPS data. We test the framework on a large user sample of 107 persons in Switzerland and interpret their travel mode choice behavior patterns based on the clustering results. This facilitates understanding people's travel mode choice behavior in multimodal transportation and how to design reasonable alternatives to private cars for more sustainable cities.

Keywords Human movement data · Travel mode choice behavior · Autocorrelation · Hierarchical clustering

1 Introduction

As one of the environmentally relevant behaviors, travel mode choice has become increasingly important with the rising social concern for the environment (Hunecke

P. Zhao (✉) · D. Bucher · H. Martin · M. Raubal
Institute of Cartography and Geoinformation, ETH Zurich,
Stefano-Franscini-Platz 5, 8093 Zurich, Switzerland
e-mail: pezhao@ethz.ch

© Springer Nature Switzerland AG 2020
P. Kyriakidis et al. (eds.), *Geospatial Technologies for Local and Regional Development*, Lecture Notes in Geoinformation and Cartography,
https://doi.org/10.1007/978-3-030-14745-7_5

et al. 2001). Specifically, traffic-related carbon dioxide (CO₂) emissions resulting from cars and taxis speed up the greenhouse effect and pose a large threat to the environment (Zhao et al. 2017a). When trying to provide sustainable travel modes, policy-makers and mobility providers are facing both challenges and opportunities. In recent years, multimodality has already been considered for a more sustainable future urban mobility, as it offers an attractive alternative to the private car through combinations of (more ecological) travel modes such as public transport, electric car (e-car) and bicycle (Klinger 2017). Especially, the proliferation of *Mobility as a Service* (MaaS) makes it more convenient for people to book a delivery service with a range of travel modes. Compared to private cars and taxis, the use of public transport and e-cars is a more environmentally friendly type of travel behavior due to their lower CO₂ emissions. Therefore, travel mode choice has always been a continuing research topic in the fields of Geographic Information Science and transportation planning.

Over the past decades, the vast majority of studies on travel mode choice have concentrated on modeling and analyzing the related influencing factors from travel surveys and questionnaires (Chen et al. 2008; Murtagh et al. 2012; An et al. 2015; Daisy et al. 2018). However, less emphasis has been given to explore individuals' travel mode choice behavior patterns. Although examining the influence factors of travel mode choice is necessary, understanding individuals' travel mode choice behavior patterns is also significant for urban transportation planning. For instance, it is meaningful to discover whether there is a large group of people who choose the car as sole travel mode or if there are people who usually choose a combination of travel modes (Heinen and Chatterjee 2015). Due to the high cost and large time requirements, travel survey data normally merely record the respondents' status quo of travel mode choice behavior, thereby ignoring the behavioral change processes. Hence, these datasets fail to uncover and analyze interpersonal and intrapersonal travel mode choice behavior patterns.

The emergence and prevalence of GPS-based human movement data (e.g. mobile phone records, taxi trajectory data) facilitates characterizing and analyzing human mobility patterns at the aggregate and individual levels (Bucher et al. 2019; Yuan and Raubal 2014; Zhao et al. 2017b; Jonietz and Bucher 2018). Undoubtedly, there are numerous studies which investigate human travel behavior patterns from movement trajectory data. For instance, Barbosa et al. (2018) reviewed research on reproducing human mobility patterns from various movement data sources in recent years, which shed light on the fundamental modeling approaches and technical methods of human mobility. However, there has not been sufficient research on exploring and understanding human travel mode choice patterns at the individual level. The objective of this work is to understand individuals' travel mode choice behavior patterns through categorizing users with similar mode choice behavior patterns based on their trajectory data. Exploring these patterns will be helpful for policy-makers to understand people's travel mode choice behavior, and design and implement more sustainable mobility strategies.

In this study, we propose a clustering-based framework for understanding individuals' travel mode choice behavior patterns. Specifically, the framework contains

three main steps: (1) construct features in terms of time series to describe individuals' travel mode choice behavior; (2) measure the similarity of different users based on autocorrelation coefficients; (3) segregate individuals into groups based on their similarity using hierarchical clustering. The reason for choosing hierarchical clustering is that researchers can define their own similarity or distance measures to generate a similarity matrix according to their research purpose and because it can be conveniently visualized using a dendrogram. The experimental dataset comes from the large-scale pilot study *SBB Green Class* conducted in Switzerland, which will be introduced in Sect. 3.

The article is organized as follows: Related work regarding travel mode choice and time series clustering are reviewed in Sect. 2. Section 3 describes the dataset used in this study. Section 4 presents the overall framework for understanding travel mode choice behavior patterns. The experimental results are shown in Sect. 5 and we discuss and conclude this research in Sect. 6.

2 Related Work

2.1 Examining Individuals' Travel Mode Choice

Travel mode choice has been investigated based on choice behavior theory for several decades (Glasser 1999). As mentioned in the introduction, the central issue of the related studies is to model and analyze the relationship between people's travel mode choices and their influencing factors. It has been demonstrated that travel mode choice is impacted by a variety of factors, such as built environment (Chen et al. 2008; Ding et al. 2017; Sun et al. 2017; Han et al. 2018), individual characteristics (Murtagh et al. 2012; Vij et al. 2013; Böcker et al. 2017), weather conditions (Böcker et al. 2013; Liu et al. 2015), or travel time and distance (An et al. 2015; Daisy et al. 2018). However, little attention has been paid to studies on detecting individuals' travel mode choice behavior patterns. Although there are several studies that investigate travelers' mode choice behavior by grouping individuals, they normally divide the travelers into groups according to socioeconomic information and neglect the behavioral change processes of travel mode choice (Ding and Zhang 2016).

With the widespread usage of smart phones and location-based services (Huang et al. 2018), it has become convenient to record individuals' daily travel activities over longer periods. The emergence of GPS-based human movement data has spurred numerous studies on individuals' travel behavior patterns from their trajectories. For instance, Shen and Cheng (2016) proposed a theoretical framework to divide users into subgroups according to their travel behavior patterns using individual trajectory data. An integrated framework was developed to analyze human mobility patterns from volunteered GPS trajectories and contextual information. Specifically, how individuals' travel mode choice depends on their residential location, age or gender (Siła-Nowicka et al. 2016). The individual travel behavior regularity was

investigated through constructing a travel behavior graph model. The graph can be used to indicate an individual's travel routes and forecast travel mode choice behavior (Liang et al. 2018). Several studies specifically considered how human movement data in combination with location-aware information technology can be used to influence mobility behavior (Froehlich et al. 2009; Bucher et al. 2016). This line of research is often driven by the desire to lower greenhouse gas emissions and make mobility more sustainable (Weiser et al. 2016). In this study, we focus on understanding individuals' travel mode choice behavior patterns, more specifically on the temporal patterns of individuals' travel duration and distance by means of each type of travel mode, which reflects the behavioral change processes of travel mode choice. Autocorrelation is employed to measure the similarity of time series, as it allows capturing regularities inherent in some behavior, and can be used to effectively calculate the similarity of short time series (Aghabozorgi et al. 2015). The background of clustering time series based on their similarity is presented in Sect. 2.2.

2.2 Clustering Time Series Based on their Similarity

Clustering time series datasets has been universally done in diverse scientific disciplines and domains. It facilitates researchers and data analysts to discover valuable information and knowledge in an unsupervised way. Measuring the similarity of time series has always been an important research question regarding time series clustering (Gunopulos and Das 2001). The traditional method to measure the similarity of time series is the Euclidean distance (ED), which regards time series as a vector across all time points and then calculates the sum of the distances between corresponding points. In recent years, a growing body of measurement methods has emerged to calculate the similarity of time series.

For time series clustering, similarity measurement approaches of time series data were summarized into four categories according to the characteristics of the time series, namely *shape-based*, *compression-based*, *model-based* and *feature-based* approaches (Aghabozorgi et al. 2015). Shape-based similarity measures are essentially used to discover similar time series in shape and time, and include Dynamic Time Warping (DTW), or Longest Common Subsequence (LCSS) (Górecki 2018). As a classical shape-based similarity measurement, DTW has been widely utilized in time series clustering. For instance, Yuan and Raubal (2012) explored dynamic urban mobility patterns from mobile phone data by measuring the similarity of time series representing the dynamic mobility patterns of different urban areas. Subsequently, a large number of improved DTW methods was proposed to measure the similarity of time series (Bankó and Abonyi 2012; Łuczak 2016; Ye et al. 2017). Compression-based similarity is applicable in short and long time series, and includes Autocorrelation or Pearson's correlation coefficient. For example, Yue et al. (2018) developed a spectral clustering framework to understand the intertwined usage of bus, metro, and taxi in urban space. Specifically, the similarity of time series that represents the

ridership patterns of mass transit modes in different urban spaces was measured by calculating autocorrelation coefficients. Model-based and feature-based measures are suitable for long time series, and include ARMA models (Xiong and Yeung 2002), Hidden Markov Models (HMM) (Dias et al. 2015), Fourier transformation (Gao et al. 2015), and wavelet transformation (Barragan et al. 2016).

Since the time series in this work are generated from daily aggregates and are thus rather short, we confine the similarity measurements to shape-based and compression-based approaches. Compared with autocorrelation, the DTW algorithm neglects the interior correlation within the time series. Considering the regularity and periodicity of individuals' travel behavior patterns, an autocorrelation approach is chosen to measure the similarity between the time series, which will be introduced in Sect. 4.3 in detail. Although autocorrelation has been employed to measure the similarity of time series, to the best of our knowledge no studies have applied it to explore individuals' travel mode choice behavior patterns.

3 Data

This study is based on the dataset from the large-scale pilot project *SBB Green Class* in Switzerland. *SBB Green Class* was carried out by the Swiss Federal Railways (SBB) and offered 139 participants a *Mobility as a Service* (MaaS) package, which included a general public transport pass, a BMW i3 electric vehicle (with a corresponding charging station at home), memberships to common car- and bikesharing programs, as well as a parking spot at a train station of the participant's choice. While the participants were primarily selected based on their geographic location, the (financial) participation preconditions lead to a bias towards middle and upper class people. As part of the pilot, the participants were asked to install a commercial application on their smart phones to track their daily movement. The recorded GPS data (approx. one GPS recording every 1–5 min, with a spatial accuracy in the order of tens of meters) were automatically segmented into trajectories and stay points by the app. Next to the raw spatial and temporal information, the data include semantic information about the travel mode of each trip and the purposes of stay points, which have been manually validated by the users themselves (the transport modes proposed for validation were identified by the commercial tracking app, but likely to be based on accelerometer data). The validated travel modes contain *airplane*, *bicycle*, *boat*, *bus*, *car*, *coach*, *e-bicycle*, *e-car*, *train*, *tram* and *walk*. While the whole project ran from January to December 2017, we here select a subset of 183 days, covering April to September 2017. In addition, selecting April 2017 as a starting date for this study allowed people at least one month's time to explore the new mobility options and settle for a regular behavior (i.e. the novelty of the MaaS offer and especially the electric car has worn off).

Since this study focuses on individuals' daily travel trips, we exclude airplane trips. Additionally, in such a practical application, GPS trajectories are not perfect due to various influencing factors, and people have gaps in their recordings (e.g.,

due to a lack of battery or voluntarily turning off the tracking device). Therefore, we filter the dataset based on the available data to ensure the trajectories are reliable. As a filtering criteria, we remove all users that either have a gap of more than 30 days duration or who recorded data on less than 150 days. This leaves us with 107 active users with mostly continuous records which are further used within this study.

4 Method

In this study, a clustering-based framework is proposed to understand individuals' travel mode choice behavior using GPS trajectory data. The overall framework consists of five steps, which are shown in Fig. 1. First, we construct features to represent individual travel mode choice behavior. The second step is to interpolate the gaps for the days of data loss. Third, we measure the similarity of individual travel mode choice behaviors based on autocorrelation. The fourth step is to divide the individuals into different groups by means of a hierarchical clustering algorithm. The last step is to conduct behavior pattern analysis based on the clustering results.

4.1 Constructing Features of Travel Mode Choice

In this section, we aim to extract a series of descriptive features of the individuals' travel mode choice behavior, namely the modal split during the same period (Jonietz et al. 2018). For instance, how long (i.e. duration) or how far (i.e. distance) someone utilizes various travel modes every day. On this basis, we construct duration-based and distance-based features respectively to depict the travel mode choice behavior for each user, which capture the temporal fluctuations of duration and distance by means of a certain travel mode during a period. In this study, one day is selected as the temporal granularity. Additionally, we only consider days were we have tracked over 70% of a users day and treat all other days as missing values.

For an individual's travel duration and distance by means of a certain type of travel mode, the features T_i and D_i can be denoted as 1×183 (days) vectors:

$$T_i = [t_i^1, t_i^2, \dots, t_i^{183}] \quad (1)$$



Fig. 1 Workflow of the framework: the raw trajectory data are processed in five consecutive steps

$$D_i = [d_i^1, d_i^2, \dots, d_i^{183}] \quad (2)$$

where $i \in (1, 2, \dots, 10)$, which corresponds to ten types of travel modes (i.e. bicycle, boat, bus, car, coach, e-bicycle, e-car, train, tram and walk). t_i^j and d_i^j represent the duration and distance of the i th type of travel mode on the j th day, respectively.

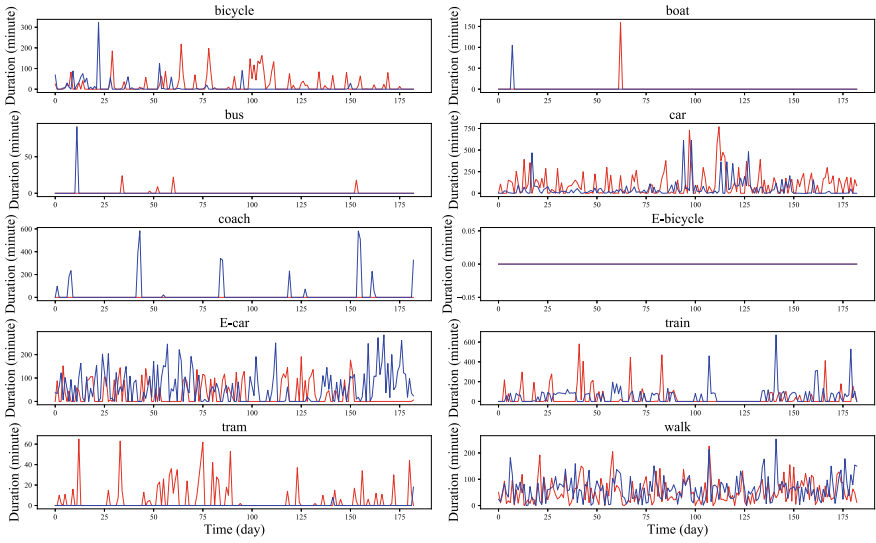
Ultimately, duration-based and distance-based features are obtained, which are expressed as 10 time series respectively. Figure 2 visualizes the constructed features of two exemplary users based on the above-mentioned 10 types of travel modes. As can be seen in the features, the two users exhibit different travel mode choice behavior patterns. The first user (red) selects multiple travel modes for daily trips, including car, e-car, tram and walk, while the second user (blue) mainly depends on e-car and walk as well as car for long-distance trips.

4.2 Interpolating the Gaps

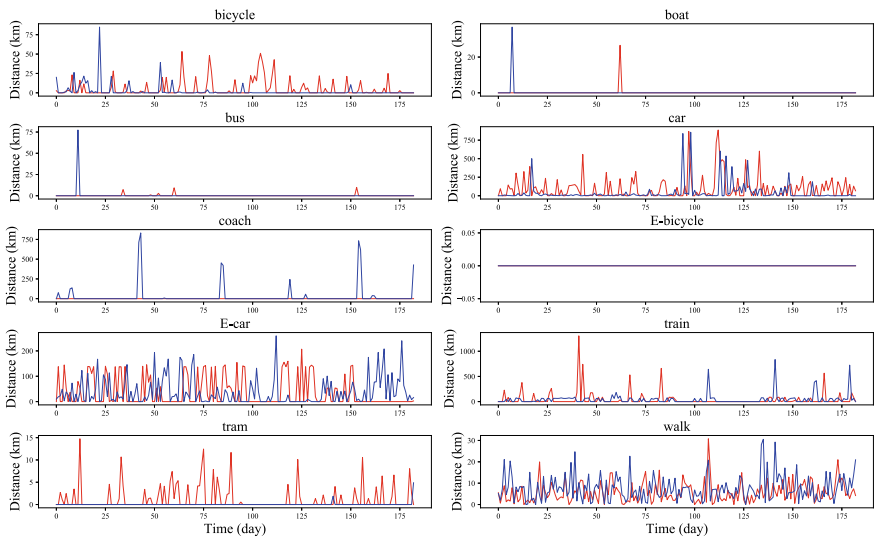
Since significant proportions of human activities occur indoors, signal loss and signal noise are prevalent in human trajectory data (Hwang et al. 2018). Additionally, trajectory data recorded with smart phone applications are also influenced by other factors (e.g. battery capacity of smart phones). Uncertainties in the individuals' trajectories make it complicated to explore their travel mode choice behavior patterns. Therefore, the features are interpolated for the days of data loss. We propose a solution to interpolate the gaps based on the previous and next workdays (or weekends). The assumption is that human mobility patterns are regular, periodic and predictable, which has been demonstrated by several related studies (Gonzalez et al. 2008; Song et al. 2010; Yuan and Raubal 2016). Considering that people's travel habits are normally stable, we choose the features of the adjacent days to impute the gaps. Concretely, the gap will be interpolated with the features of two workdays adjacent to it if the trajectories are missing on a workday. Likewise, the data loss on weekends is processed in a similar way.

4.3 Measuring Individuals' Similarity Based on Autocorrelation

Based on the aforementioned features, the goal of this section is to measure the similarity between different users. Since each feature is in the form of a time series, we calculate the similarities of the time series based on their autocorrelation coefficients (AC) (D'Urso and Maharaj 2009). Compared to similarity measurement methods based on the shape of the time series, autocorrelation considers interior correlation characteristics of the time series. Given a set of time series $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_K\}$,



(a) duration



(b) distance

Fig. 2 Temporal variations of the duration and distance by means of various travel modes for two exemplary users (red and blue)

as shown in Eq. (3), the autocorrelation coefficient of the k th time series at time lag r can be expressed as Eq. (4):

$$\mathbf{X} = \begin{pmatrix} x_{11} & \dots & x_{k1} & \dots & x_{K1} \\ \vdots & & \vdots & & \vdots \\ x_{1t} & \dots & x_{kt} & \dots & x_{Kt} \\ \vdots & & \vdots & & \vdots \\ x_{1T} & \dots & x_{kT} & \dots & x_{KT} \end{pmatrix} \quad (3)$$

$$\hat{\rho}_{kr} = \frac{\sum_{t=r+1}^T (x_{kt} - \bar{x}_k)(x_{k(t-r)} - \bar{x}_k)}{\sum_{t=1}^T (x_{kt} - \bar{x}_k)^2} \quad (4)$$

where K is the number of time series, T is the length of one time series, $\mathbf{x}_k = \{x_{kt} : t = 1, 2, \dots, T\}$ represents the k th ($k = 1, 2, \dots, K$) time series, x_{kt} stands for the t th observation value of the k th time series, and \bar{x}_k is the mean of the k th time series.

As we want to classify people not solely on the regularity of their transport mode choices, but also by taking into account the relative number of times they used a certain transport mode, we introduce a user-specific weighting of each autocorrelation vector $\hat{\rho}_k$. Namely, as each autocorrelation vector is computed from either the daily distance or duration of a single mode of transport, we multiply it by the share of this mode's distance or duration over the whole study period for each user ($w_k = d_k / \sum_{k=1}^{10} d_k$, where d_k is either the distance or duration of transport mode k , depending on how the respective autocorrelation vector was computed). To clarify, let us assume there are two users: one never uses the car, while the other uses the car every day to go to work. Simply looking at the autocorrelation would classify these users into the same class, as their time series are perfectly autocorrelated. Introducing user-specific weights for the autocorrelation, essentially scales the autocorrelation values of the first user to zero, thus increasing the likelihood that this user gets into another cluster.

Given two sets of time series for two users' travel duration by use of various travel modes and the determined time lag R , two sets of autocorrelation vectors can be calculated accordingly on the basis of Eq. (4). We assume the autocorrelation vectors $\hat{\rho}_{si} = (\hat{\rho}_{si1}, \dots, \hat{\rho}_{siR})$ and $\hat{\rho}_{ti} = (\hat{\rho}_{ti1}, \dots, \hat{\rho}_{tiR})$ correspond to the i th travel mode for users s and t respectively, and the weights of them are $\mathbf{w}_s = (w_{s1}, \dots, w_{s10})$ and $\mathbf{w}_t = (w_{t1}, \dots, w_{t10})$. The similarity of two users is measured based on the autocorrelation vectors and weights, which is denoted by the follow formula:

$$d_{st}^2 = \sum_{i=1}^{10} \sum_{r=1}^R (\hat{\rho}_{sir} \cdot w_{si} - \hat{\rho}_{tir} \cdot w_{ti})^2 \quad (5)$$

where $i = 1, 2, \dots, 10$ stands for the i th travel mode, $r = 1, 2, \dots, R$ are the time lags for the time series.

4.4 Clustering the Time Series

We measure the similarities of individuals by calculating the distance of their corresponding weighted autocorrelation vectors at different time lags with Eq. 5 and obtain the similarity matrix. On the basis of the similarity matrix, the individuals can be categorized into different groups based on their travel mode choice behavior changes. In this work, we choose a hierarchical clustering algorithm to process the similarity matrix (Gower and Ross 1969). Hierarchical clustering algorithms attempt to construct a hierarchy of clusters through a bottom-up (i.e. agglomerative) or top-down strategy (i.e. divisive). The output of the algorithm is a dendrogram, which intuitively displays the hierarchical relationships between the clusters. One of its major advantages is that hierarchical clustering is flexible for inputting a similarity matrix generated by various similarity or distance measures according to the research purpose. Therefore, it has been widely used in human mobility analysis (Shen and Cheng 2016; Wang et al. 2018).

The number of clusters to be generated in this study is determined by the Calinski-Harabasz Index (CH-index) (Caliński and Harabasz 1974), which evaluates how well the dataset is separated quantitatively. Normally, the location of the maximum CH-index corresponds to the optimal number of clusters. Before calculating the CH-index, two basic elements, namely SSW and SSB (sum of squares within resp. between the clusters) need to be calculated, which can be used to quantify the overall within-cluster and between-cluster variances. The CH-index is a ratio based on them. Given a set $X = \{x_1, x_2, \dots, x_N\}$, representing a dataset with N data points, $\bar{X} = \sum_{i=1}^N x_i / N$ is the center of the whole dataset. Let us denote the centroids of clusters $\{C_1, C_2, \dots, C_M\}$ as $C = \{c_1, c_2, \dots, c_M\}$, with M being the number of clusters and n_i being the number of elements in cluster C_i . SSW , SSB and CH are then computed as follows:

$$SSW = \sum_{i=1}^N \|x_i - C_{p_i}\|^2 \quad (6)$$

$$SSB = \sum_{i=1}^M n_i \|c_i - \bar{X}\|^2 \quad (7)$$

$$CH = \frac{SSB / (M - 1)}{SSW / (N - M)} \quad (8)$$

where x_i represents the i th point, C_{p_i} indicates the centroid of the p th cluster and x_i is in the p th cluster.

5 Results

5.1 *Measuring the Similarity of Individuals*

This section aims at measuring the individuals' similarity with autocorrelation based on the constructed time series that represent their travel mode choice behavior, which will provide input for the clustering analysis. The autocorrelation coefficients at different time lags constitute the autocorrelation vector of a time series. Theoretically, the time lag r is between 1 and $T - 1$ (where T is the number of days in the time series). Considering the specific characteristics of the constructed features (e.g. regularity and periodicity), we examine the differences of their autocorrelation coefficients at different time lag limits. We select all users' travel duration and distance by train as an example. The variations of autocorrelation coefficients for duration and distance at different time lags are shown in Fig. 3. To observe the variations of the autocorrelation coefficients more clearly, Fig. 3 (c) and (d) visualize the first 50 autocorrelation coefficients. It can be seen that the autocorrelation coefficients display regular weekly patterns. According to the recommendation by (Box et al. 2015) that the number of lags up to about a quarter of the time series length is sufficient to assess the dependence structure of the time series, this study examines the autocorrelation coefficients up to lags 7, 14, 21, and 28 for the calculations of similarity.

Based on the selected maximum time lag, similarity matrices for both duration and distance can be calculated using the procedure introduced in Sect. 4.3. Here, the two similarity matrices are calculated based on the autocorrelation coefficients up to lags 28 to reflect the overall similarity of individuals' travel mode choice behavior in terms of duration and distance. In addition, we further calculate the correlation coefficient of the two matrices. The correlation coefficient reaches 0.86, which implies that travel duration and distance are comparatively consistent in the depiction of individuals' travel mode choice behavior.

5.2 *Detecting Travel Mode Choice Behavior Patterns*

On the basis of similarity matrices of duration-based and distance-based features, the goal of this section is to segregate all individuals into groups. The individuals are divided into subgroups in the form of a hierarchical tree. The hierarchical tree can be cut at certain predetermined locations to divide the whole dataset into several groups. We utilize the CH-index to determine the number of clusters, which evaluates the clustering validity based on the average between- and within- cluster sum of squares. Based on the similarity matrices, Fig. 4 presents the relations between the number of clusters and CH-index for different time lags. Note that the optimal number of clusters is different for the different number of time lags. According to the optimal number of clusters determined by Fig. 4, Table 1 displays the clustering results for different time lags. Specifically, the clusters with one or two users are regarded as

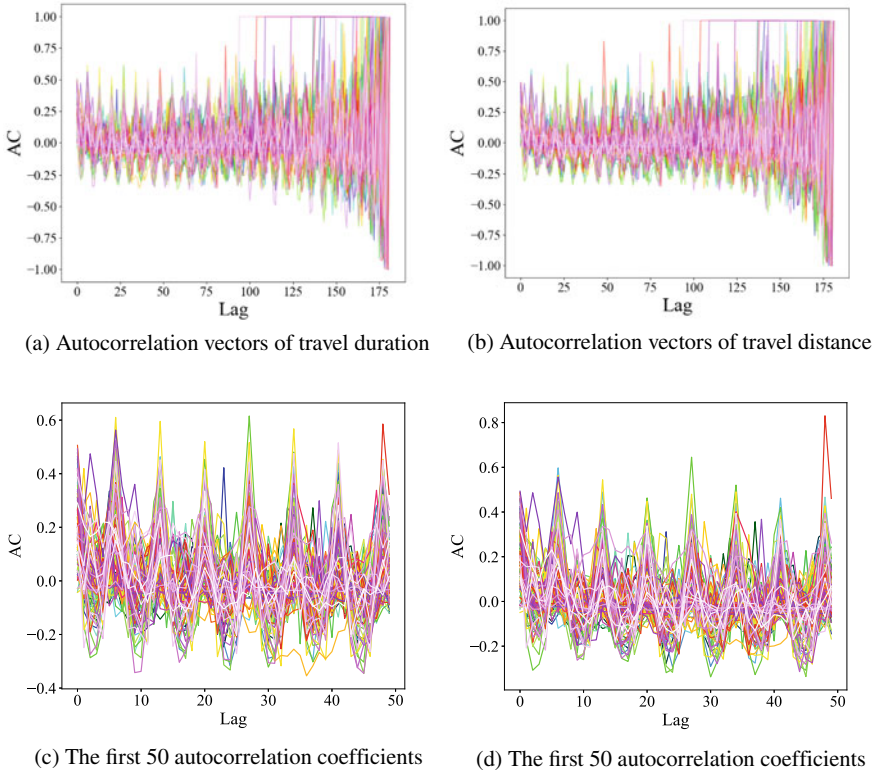


Fig. 3 Variations of the autocorrelation coefficients at different time lags for all users

outliers. Next, we analyze the users’ travel mode choice behavior patterns at the aggregate and individual levels respectively based on the fourth clustering results (i.e. the number of lags is 28).

First, we analyze the travel mode choice patterns at the aggregate level. Specifically, the proportions of travel duration and distance by means of various travel modes are calculated for each cluster, as shown in Table 2. It can be observed that the individuals in each cluster show different transport mode use patterns. For example, car and train are chosen as the main travel modes, and e-car and walk as secondary travel modes in cluster 1. For cluster 2, train occupies a high proportion as the main travel mode, while car, e-car and walk occupy comparatively low proportions as secondary travel modes. Meanwhile, note that the two clusters based on duration show the same main travel modes as those based on distance, namely car and train, train, as well as car. It also demonstrates that travel duration and distance are consistent in describing individuals’ travel mode choice behavior. However, the difference between them is also worth noting. For instance, walk does not appear as secondary travel mode in the clusters based on distance, which is also in accordance with individuals’ daily trips. After all, walk is normally selected for short-distance trips.

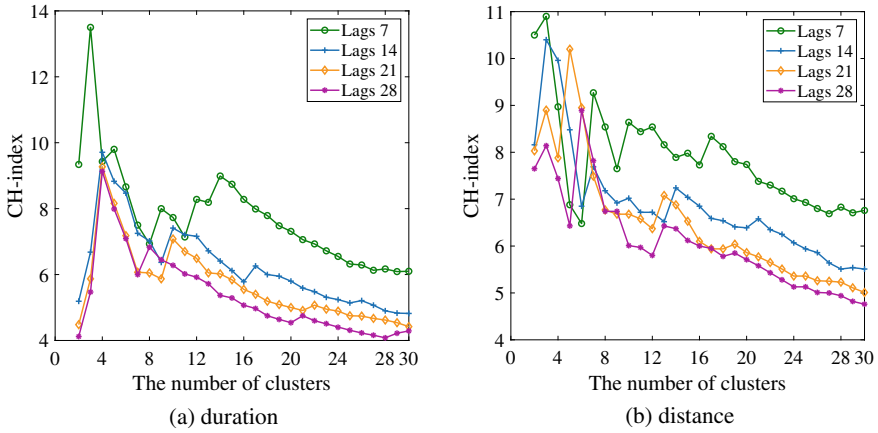


Fig. 4 Relation graph of the number of clusters and CH-index

Table 1 Results of the clustering for different time lags

	Num(lags) = 7		Num(lags) = 14		Num(lags) = 21		Num(lags) = 28	
	Cluster	Num	Cluster	Num	Cluster	Num	Cluster	Num
Duration	1	100	1	100	1	100	1	100
	2	6	2	5	2	5	2	5
	3	1	3	1	3	1	3	1
			4	1	4	1	4	1
Distance	1	104	1	104	1	94	1	93
	2	2	2	2	2	9	2	9
	3	1	3	1	3	2	3	1
					4	1	4	2
					5	1	5	1
							6	1

Second, travel mode choice behavior patterns at the individual level are investigated. Concretely, five human mobility indicators are selected to explore individuals' travel mode choice behavior patterns, including travel duration and distance based on four travel modes (i.e. car, e-car, train and walk) as well as carbon dioxide (CO₂) emissions. These emissions were computed by multiplying the distances covered with each means of transport by a mode-specific constant as given by the *Swiss platform for mobility management tools (mobitool)* (Tuchschmid and Halder 2010). The nine resulting indicators refer to the mean of the corresponding observation values in the selected period for each user. In addition, we normalize these indicators to ensure that they fall between 0 and 1 in order to conveniently compare them in the same figure. To understand the distribution of indicators for the users in each cluster, we visualize their distribution using a boxplot in Fig. 5. Looking at Fig. 5a, it can

Table 2 The shares of travel modes for different clusters

Travel mode	Duration		Distance	
	Cluster 1	Cluster 2	Cluster 1	Cluster 2
Bicycle	0.038	0.021	0.017	0.008
Boat	0.009	0.015	0.003	0.001
Bus	0.014	0.015	0.008	0.005
Car	0.314	0.146	0.373	0.208
Coach	0.003	0.005	0.004	0.000
e-bicycle	0.000	0.000	0.000	0.000
e-car	0.173	0.129	0.210	0.131
Train	0.252	0.488	0.351	0.622
Tram	0.008	0.008	0.006	0.003
Walk	0.189	0.173	0.028	0.022

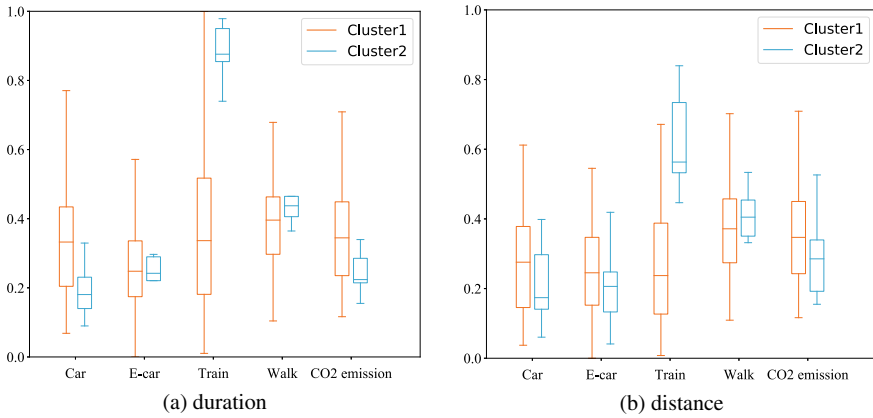


Fig. 5 Boxplots of five mobility indicators in different clusters

be seen that the users in clusters 1 spend more time by car for their trips than those of cluster 2 as a whole. Similarly, we can conclude that the users in cluster 2 spend more time by train for their trips than those of clusters 1, which results from train as the sole main travel mode for the users in cluster 2. It is noteworthy that the travel time by e-car and walk is not very different for the users in the two clusters, which is due to the fact that e-car and walk are both regarded as secondary travel modes. Additionally, another remarkable phenomenon is seen in the comparison of CO₂ emissions among the two clusters. High CO₂ emissions of clusters 1 indicate that car is probably still the main source of transport-related carbon dioxide emissions. Public transport and green travel modes (e.g. e-car) are efficient alternatives for the

reduction of CO₂ emissions and for sustainable urban development. Similar analyses can be conducted to explain the travel mode choice behavior patterns for the distance-based indicators.

6 Conclusions and Future Work

The availability of long-term human movement data with semantic information (e.g. travel mode) enables us to investigate individuals' travel mode choice behavior patterns. Specifically, behavioral change processes cannot be considered in traditional travel mode choice studies due to a lack of advanced data collection methods. In this study, we propose a clustering-based framework to understand individuals' travel mode choice behavior patterns by segregating them into groups that exhibit similar travel mode choice behavior, based on a human trajectory dataset with semantic information. First, we construct duration-based and distance-based features in the form of time series to depict individuals' travel mode choice behavior. Next, we propose a weighted autocorrelation method to measure the similarity of individuals. Finally, a hierarchical clustering algorithm is employed to divide the individuals into groups based on the similarity matrix. For a case study in Switzerland, two clusters of individuals are detected and interpreted from the perspective of travel mode choice behavior patterns at the aggregate and individual levels. Our contributions facilitate understanding people's travel mode choice behavior in multimodal transportation and how to design reasonable alternatives to private cars for more sustainable cities. In addition, dividing the individuals into different groups based on their travel mode choice behavior will also help policymakers design and provide personalized travel mode recommendation services for different user groups.

However, there are several limitations in the current study, which could be regarded as directions for future work. First, the interpolation of data loss is simply conducted based on the historical movement data. Especially, only the data of two adjacent days are used, the effect of which requires to be considered. Hence, new methods and techniques of data gaps imputation should be investigated to impute gaps in the trajectory dataset. Second, the current hierarchical clustering algorithm has limited ability working on high-dimension datasets. New advanced high-dimensional clustering methods would be more appropriate for this study. Last but not least, this study only analyzes the individuals' travel mode choice behavior patterns based on the clustering results. Although it can reflect the difference between individuals, it would be more meaningful to explore the cross influence of socio-demographic and urban environment factors on the patterns.

Acknowledgements This research was supported by the Swiss Data Science Center (SDSC), by the Swiss Innovation Agency Innosuisse within the Swiss Competence Center for Energy Research (SCCER) Mobility and by the Swiss Federal Railways SBB.

References

- Aghabozorgi S, Shirkhorshidi AS, Wah TY (2015) Time-series clustering—a decade review. *Inf Syst* 53:16–38
- An S, Wang Z, Cui J (2015) Integrating regret psychology to travel mode choice for a transit-oriented evacuation strategy. *Sustainability* 7(7):8116–8131
- Bankó Z, Abonyi J (2012) Correlation based dynamic time warping of multivariate time series. *Expert Syst Appl* 39(17):12814–12823
- Barbosa H, Barthelemy M, Ghoshal G, James CR, Lenormand M, Louail T, Menezes R, Ramasco JJ, Simini F, Tomasini M (2018) Human mobility: models and applications. *Phys Rep*
- Barragan JF, Fontes CH, Embiruçu M (2016) A wavelet-based clustering of multivariate time series using a multiscale spca approach. *Comput Ind Eng* 95:144–155
- Böcker L, Prillwitz J, Dijst M (2013) Climate change impacts on mode choices and travelled distances: a comparison of present with 2050 weather conditions for the randstad holland. *J Transp Geogr* 28:176–185
- Böcker L, van Amen P, Helbich M (2017) Elderly travel frequencies and transport mode choices in Greater Rotterdam, the Netherlands. *Transportation* 44(4):831–852
- Box GE, Jenkins GM, Reinsel GC, Ljung GM (2015) *Time series analysis: forecasting and control*. John Wiley, New Jersey
- Bucher D, Cellina F, Mangili F, Raubal M, Rudel R, Rizzoli AE, Elabed O (2016) Exploiting fitness apps for sustainable mobility—challenges deploying the Goeco! app. *ICT for sustainability (ICT4S)*
- Bucher D, Mangili F, Cellina F, Bonesana C, Jonietz D, Raubal M (2019) From location tracking to personalized eco-feedback: a framework for geographic information collection, processing and visualization to promote sustainable mobility behaviors. *Travel Behav Soc* 14:43–56
- Calinski T, Harabasz J (1974) A dendrite method for cluster analysis. *Commun Stat-Theory Methods* 3(1):1–27
- Chen C, Gong H, Paaswell R (2008) Role of the built environment on mode choice decisions: additional evidence on the impact of density. *Transportation* 35(3):285–299
- Daisy NS, Millward H, Liu L (2018) Trip chaining and tour mode choice of non-workers grouped by daily activity patterns. *J Transp Geogr* 69:150–162
- Dias JG, Vermunt JK, Ramos S (2015) Clustering financial time series: new insights from an extended hidden markov model. *Eur J Oper Res* 243(3):852–864
- Ding C, Wang D, Liu C, Zhang Y, Yang J (2017) Exploring the influence of built environment on travel mode choice considering the mediating effects of car ownership and travel distance. *Transp Res Part A: Policy Pract* 100:65–80
- Ding L, Zhang N (2016) A travel mode choice model using individual grouping based on cluster analysis. *Procedia Eng* 137:786–795
- D’Urso P, Maharaj EA (2009) Autocorrelation-based fuzzy clustering of time series. *Fuzzy Sets Syst* 160(24):3565–3589
- Froehlich J, Dillahunt T, Klasnja P, Mankoff J, Consolvo S, Harrison B, Landay JA (2009) Ubigreen: investigating a mobile tool for tracking and supporting green transportation habits. In: *Proceedings of the SIGCHI conference on human factors in computing systems, ACM*, pp 1043–1052
- Gao Z-K, Yang Y-X, Fang P-C, Jin N-D, Xia C-Y, Hu L-D (2015) Multi-frequency complex network from time series for uncovering oil-water flow structure. *Sci Rep* 5:8222
- Glasser W (1999) *Choice theory: a new psychology of personal freedom*. Harper Perennial, New York
- Gonzalez MC, Hidalgo CA, Barabasi A-L (2008) Understanding individual human mobility patterns. *Nature* 453(7196):779
- Górecki T (2018) Classification of time series using combination of DTW and LCSS dissimilarity measures. *Commun Stat-Simul Comput* 47(1):263–276
- Gower JC, Ross GJ (1969) Minimum spanning trees and single linkage cluster analysis. *Appl Stat* 54–64

- Gunopulos D, Das G (2001) Time series similarity measures and time series indexing. *Acm Sigmod Record*, vol 30, ACM, p 624
- Han Y, Li W, Wei S, Zhang T (2018) Research on passenger's travel mode choice behavior waiting at bus station based on sem-logit integration model. *Sustainability* 10(6):1996
- Heinen E, Chatterjee K (2015) The same mode again? an exploration of mode choice variability in great britain using the national travel survey. *Transp Res Part A: Policy Pract* 78:266–282
- Huang H, Gartner G, Krisp JM, Raubal M, de Weghe NV (2018) Location based services: ongoing evolution and research agenda. *J Locat Based Serv* 12(2):63–93
- Hunecke M, Blöbaum A, Matthies E, Höger R (2001) Responsibility and environment: ecological norm orientation and external factors in the domain of travel mode choice behavior. *Environ Behav* 33(6):830–852
- Hwang S, VanDeMark C, Dhatt N, Yalla SV, Crews RT (2018) Segmenting human trajectory data by movement states while addressing signal loss and signal noise. *Int J Geogr Inf Sci* 32(7):1391–1412
- Jonietz D, Bucher D (2018) Continuous trajectory pattern mining for mobility behaviour change detection. In: *LBS 2018: 14th international conference on location based services*. Springer, pp 211–230
- Jonietz D, Bucher D, Martin H, Raubal M (2018) Identifying and interpreting clusters of persons with similar mobility behaviour change processes. In: Mansourian A, Pilesjö P, Harrie L, van Lammeren R (eds) *AGILE 2018—geospatial technologies for all*. Springer International Publishing, Cham, pp 291–307
- Klinger T (2017) Moving from monomodality to multimodality? changes in mode choice of new residents. *Transp Res Part A: Policy Pract* 104:221–237
- Liang Q, Weng J, Zhou W, Santamaria SB, Ma J, Rong J (2018) Individual travel behavior modeling of public transport passenger based on graph construction. *J Adv Transp* 2018
- Liu C, Susilo YO, Karlström A (2015) The influence of weather characteristics variability on individual's travel mode choice in different seasons and regions in Sweden. *Transp Policy* 41:147–158
- Łuczak M (2016) Hierarchical clustering of time series data with parametric derivative dynamic time warping. *Expert Syst Appl* 62:116–130
- Murtagh N, Gatersleben B, Uzzell D (2012) Multiple identities and travel mode choice for regular journeys. *Transp Res Part F: Traffic Psychol Behav* 15(5):514–524
- Shen J, Cheng T (2016) A framework for identifying activity groups from individual space-time profiles. *Int J Geogr Inf Sci* 30(9):1785–1805
- Siła-Nowicka K, Vandrol J, Oshan T, Long JA, Demšar U, Fotheringham AS (2016) Analysis of human mobility patterns from gps trajectories and contextual information. *Int J Geogr Inf Sci* 30(5):881–906
- Song C, Qu Z, Blumm N, Barabási A-L (2010) Limits of predictability in human mobility. *Science* 327(5968):1018–1021
- Sun B, Ermagun A, Dan B (2017) Built environmental impacts on commuting mode choice and distance: evidence from Shanghai. *Transp Res Part D: Transp Environ* 52:441–453
- Tuchschmid M, Halder M (2010) *mobitool—grundlagenbericht: Hintergrund. Methodik & Emissionsfaktoren*
- Vij A, Carrel A, Walker JL (2013) Incorporating the influence of latent modal preferences on travel mode choice behavior. *Transp Res Part A: Policy Pract* 54:164–178
- Wang Y, Qin K, Chen Y, Zhao P (2018) Detecting anomalous trajectories and behavior patterns using hierarchical clustering from taxi GPS data. *ISPRS Int J Geo-Inf* 7(1):25
- Weiser P, Scheider S, Bucher D, Kiefer P, Raubal M (2016) Towards sustainable mobility behavior: research challenges for location-aware information and communication technology. *GeoInformatica* 20(2):213–239
- Xiong Y, Yeung D-Y (2002) Mixtures of arma models for model-based time series clustering. In: *2002 IEEE International Conference on Data Mining, 2002. ICDM 2003. Proceedings, IEEE*, pp 717–720

- Ye Y, Niu C, Jiang J, Ge B, Yang K (2017) A shape based similarity measure for time series classification with weighted dynamic time warping algorithm. In: 4th International conference on information science and control engineering (ICISCE), 2017, IEEE, pp 104–109
- Yuan Y, Raubal M (2012) A framework for spatio-temporal clustering from mobile phone data. Workshop on complex data mining in a geospatial context proceedings at AGILE 2012. Association of Geographic Information Laboratories for Europe (AGILE), pp 22–26
- Yuan Y, Raubal M (2014) Measuring similarity of mobile phone user trajectories—a spatio-temporal edit distance method. *Int J Geogr Inf Sci* 28(3):496–520
- Yuan Y, Raubal M (2016) Analyzing the distribution of human activity space from mobile phone usage: an individual and urban-oriented study. *Int J Geogr Inf Sci* 30(8):1594–1621
- Yue M, Kang C, Andris C, Qin K, Liu Y, Meng Q (2018) Understanding the interplay between bus, metro, and cab ridership dynamics in shenzhen, China. *Trans GIS* 22(3):855–871
- Zhao P, Kwan M-P, Qin K (2017a) Uncovering the spatiotemporal patterns of CO2 emissions by taxis based on individuals' daily travel. *J Transp Geogr* 62:122–135
- Zhao P, Qin K, Ye X, Wang Y, Chen Y (2017b) A trajectory clustering approach based on decision graph and data field for detecting hotspots. *Int J Geogr Inf Sci* 31(6):1101–1127

Classification of Urban and Rural Routes Based on Motorcycle Riding Behaviour



Gerhard Navratil, Ioannis Giannopoulos and Gilbert Kotzbek

Abstract A basic problem in navigation is the selection of a suitable route. This requires a determination of costs or suitability. There are approaches for many standard situations, e.g., the shortest route for pedestrians, the fastest route for cars, a physically possible and legal route for trucks, or the safest route for bicycle riders. However, not much research has been done yet for motorcycle riders. Published approaches rely on interpretation of geometry, interviews, or user feedback. None of these approaches is precise and scalable. Since modern motorcycles have an increasing number of internal sensors (e.g., lean angle sensors for curve ABS), they could provide the data required for a classification of route segments. The combination with a navigational device allows to georeference the data and thus attach riding characteristics to a specific road segment. This work sketches the classification concept and presents data from a real-driving experiment using an external IMU.

Keywords Routing · Inertial measurement unit · Motorcycle · Classification · Navigation

1 Introduction

The calculation of an optimal route requires a criterion to determine the costs of each segment. Geometrical distance, the travel time, the energy consumption, or the financial costs are often applied. Any of the standard algorithms can then produce an optimal path. In other situations, this strategy does not directly work, e.g., when

G. Navratil (✉) · I. Giannopoulos · G. Kotzbek
Department for Geodesy and Geoinformation, Vienna University of Technology, Gusshausstr.
27-29, 1040 Vienna, Austria
e-mail: gerhard.navratil@geo.tuwien.ac.at

I. Giannopoulos
e-mail: igiannopoulos@geo.tuwien.ac.at

G. Kotzbek
e-mail: gilbert.kotzbek@geo.tuwien.ac.at

© Springer Nature Switzerland AG 2020
P. Kyriakidis et al. (eds.), *Geospatial Technologies for Local and Regional Development*, Lecture Notes in Geoinformation and Cartography,
https://doi.org/10.1007/978-3-030-14745-7_6

looking for a scenic route (compare Hochmair and Navratil 2008) because two different properties need to be balanced. In this case, scenic routes should be preferred over other routes of the same length. A solution is to either reduce the costs of scenic segments or increase them for unattractive segments. A multiplier to the costs would have this effect. A first attempt to do this for motorcycle routing was published in 2012 based on the utilization of the routes' geometrical properties assuming that riding a winding road is more attractive and safer than driving along a straight road (Navratil 2012). The conclusions mention a number of other parameters for suitability like grip level, legal speed limit, and traffic. This shows that the determination of optimal routes for motorcycle riders is more difficult than for other modes of transportation. More parameters affect the suitability of a segment than in the case of cars, bicycles, or pedestrians. Since motorcycle riders do not enjoy the same protection by technology like car drivers or by the lawmaker like bicyclists or pedestrians, rider concentration is an important topic. Continuously changing conditions keep the concentration at a suitable level and provide more fun than monotonous situations (e.g., on highways) or abruptly occurring hazards (like in dense urban traffic). The collection of the parameters listed by Navratil (2012) is theoretically possible. However, the assessment of the resulting attractiveness might be problematic since many different aspects need to be addressed. Klingspor (2018) avoids this problem by utilizing a different approach based on tracks suggested on motorcycle portals to assess the attractiveness of routes. The disadvantage of this approach is that the tracks are not stored as a set of edges but as a sequence of points between which optimal parts need to be calculated. It is unclear if the resulting set of edges belongs to the suggested route.

An important factor is the determination if the road segment is in an urban or a rural area. Urban areas tend to be connected to more hazards (e.g., tram tracks), more traffic (including bicycle riders and pedestrians), more stops (cause by stop signs and traffic lights), and straight lines connecting intersections. Driving speed seems to be an obvious factor differentiating between rural and urban segments. However, analysis of floating car data shows that up to 40% of the drivers are speeding in urban areas (Partusch et al. 2014). A reliable information for this separation is the current vehicle position. However, this information is not always available in the necessary absolute quality. Being able to determine different types of road segments could lead to improved navigational services. An increasing number of riders is using navigation services in unknown and even in familiar environments. These systems may be vehicle fixed or on a smartphone. The solution determined by the system could be improved by a road segment classification to select routes that are safer and more fun to drive without significantly lengthening the route.

One development of the last decade provides a basis to implement a new approach for measuring and quantifying attractiveness: Sensors in vehicles. Vehicles including motorcycles contain an increasing number of sensors to improve driving safety and efficiently manage the engine and auxiliary systems. Given sensory access restrictions of the bike used for this work, we focused on the assessment and analysis of rotation and acceleration information retrieved through an external inertial measure-

ment unit (IMU). This was done independently of location to provide a strategy also applicable in situations where the absolute position is not available.

The remainder of the paper is structured as follows: Sect. 2 provides the basis by inspecting the necessary sensors for modern motorcycles, Sect. 3 describes general expectations of what kind of information the data could unveil, Sect. 4 describes the test setup, and Sect. 5 shows the results of the data analysis. Section 6 discusses the results from the experiment, points out some further research directions, and concludes the paper.

2 Sensors in Modern Motorcycles

Since modern motorcycles (like all other vehicles driven by combustion engines) need to comply with increasing emission restrictions and rider safety becomes a strong sales argument, the sensors required for these tasks will also be present in future motorcycles. Currently, the data produced by the sensors are only available inside the system of the motorcycle and manufacturer-specific equipment is required to read and analyse them. This is comparable to the situation of geographical information systems in the early 1990ies before the foundation of the Open Geospatial Consortium led to open standards for a free exchange of data without loss of semantic information. As soon as a standardized interface for data access for motorcycles is implemented, new applications using these data can be developed. One of these could be a navigation service for motorcycle riders. However, the development of such a service requires information about the type of data collected by the sensors and a realistic assessment of suitable sensors.

Modern motorcycles depend on a large number of sensors. Some are necessary for driving assistant systems like anti-blocking system (ABS), traction control (TC), or quickshifter, others are necessary to let the engine operate in the optimal way (compare, e.g., Seiniger et al. 2008; De Filippi et al. 2011). Thus, the motorcycle is constantly monitoring data like

- revolutions per minute of engine,
- revolutions per minute of the wheels,
- speed,
- gear,
- gas handle, and
- breaking power.

Revolutions per minute for front and back wheel are required for TC and ABS. However, while the TC only needs to compare the results for front and back wheel and act if the back wheel is spinning much fast than the front wheel, an ABS must control a much more difficult situation. ABS shall prevent locking the wheel during deceleration. This may happen on either, the front or the back wheel, and both may be undesirable. It is also more complicated than with four-wheeled vehicles because motorcycles lean into curves and then the tire must cope with lateral forces as well.

Therefore, modern motorcycle ABS include lean angle sensors. Some or all of these data may be stored in the electronic system of the motorcycle to be used in case of malfunction or standard maintenance.

3 Expectations from Sensor Readings

Riding behaviour depends on a number of factors. Some are personal (riding skills, mood, time pressure, or cautiousness), some are based on the motorcycle (power-to-weight ratio, mounted tyres, possible lean angles, or agility), and affected by the road and traffic situation. Road and traffic situation is interesting information for a navigation service. Some aspects on the situation can be derived from various sources. The radius of turns can be derived from large-scale maps and the speed limit is typically published by the authorities either as a general rule (city road, rural road, highway) or in form of differing speed limits for specific sections. Weather forecast, for example, provides a general idea on temperature and precipitation. The same is true for traffic information. There are two types of online traffic flow information sources: Stationary sensors and floating car data. Stationary sensors provide accurate observations but only for a limited area. However, a suitable selection of locations for stationary sensors can still provide a reasonable overview on the traffic situation. Floating car data provide valuable information on the driving speed of vehicles (compare Partusch et al. 2014) but their use is severely restricted by privacy regulations in many countries. Nevertheless, there are several aspects that cannot be covered by these sources, e.g., local variations in weather conditions, local risks due to past weather events (slippery roads) as well as seasonal risks of similar nature. All of these aspects can influence the standard grip level of the tarmac. Thus, these aspects should influence the riding behaviour and should therefore be visible in the data. Some expectations could be formulated as follows:

- Lean angles decrease in wet conditions.
- Lean angles increase with the road grip level.
- Dense traffic in the direction of driving will lead to overtake manoeuvres that require lane changes, strong acceleration and possibly quick deceleration.
- Dense traffic in both directions will lead to more monotonous driving based on the average speed of cars and trucks.

Some of these data, like the lean angle, are available directly for the motorcycle. Others may not be observed directly but can be computed by other means. Acceleration or deceleration, for example can be derived from speed changes or changes in the engine rotations per minute.



Fig. 1 XSENS sensor and mounting on the motorcycle

4 Test Setup and Test Implementation

The motorcycle utilized for in the experiment was a KTM Duke 790.¹ As an IMU, the XSENS MTi²—miniature gyro-enhanced Attitude and Heading Reference Sensor was employed (see Fig. 1 left). It provides 3D orientation, 3D acceleration, and 3D rate of turn. The earth’s magnetic field provides a reference frame for the orientation of the sensor’s x-axis. Since the sensor was mounted such that:

- The x-axis is parallel to the wheelbase with the positive direction back.
- The y-axis is perpendicular to the x-axis pointing to the left side of the motorcycle.
- The z-axis points from the bottom of the motorcycle up

The angular readings provide roll, pitch and yaw. Roll is the rotation on the x-axis, i.e., the lean angle of the motorcycle. Pitch is the rotation around the y-axis and depends on acceleration and deceleration of the motorcycle and if the road is ascending or descending. Yaw is the rotation around the z-axis, and describes the direction of driving (the orientation).

The IMU itself has no storage capability and thus a connection by cable to the USB-port of a computer is required. The producer of the sensor also provides Windows-based software to read and store the sensor data. A Microsoft Surface Book was used to collect the data during the tests. Mounting the sensor and placing the computer was challenging since a motorcycle does not have much room for attachments. A tailbag attached to the pillion seat provided a solution. The sensor was fixed to the bottom of the tailbag using duct tape and the computer was wrapped in air bubble film to avoid damage. In practice, the chosen setting was sub optimal:

- The rider needs to unmount the motorcycle and open the tailbag to access the computer. This prevented a fine resolution in the segmentation of the track.
- The mounting was not strong enough to cope with all bumps occurring during the trips. As a result, the sensor moved slightly and some of the reading seem to have

¹<https://www.ktm.com/at/naked/790-duke/?color=ORANGE#productHeader> for technical details.

²http://opportunity-project.eu/system/files/MTi_and_MTx_User_Manual_and_Technical_Documentation.pdf.

large errors (lean angles of more than 60° on either side, for example). In principle, all erroneous data would have to be eliminated. Unfortunately, there is no possible criterion to determine, when the sensor was back in the original position. Thus, the data were not excluded from the data but small sections of the data were evaluated and compared to the full interval.

The sampling rate of the sensor was set to 10 Hz because acceleration in urban areas rarely exceeds 1 s due to the speed limits. Lower sampling rates would not allow to detect shorter periods of acceleration and these may occur in urban traffic situations (compare the Nyquist–Shannon sampling theorem: Shannon 1949).

Two different routes were selected for the experiment on two consecutive days. One route contains a large section that is driven in both directions (compare Fig. 2). This is not a problem since there is an essential difference between driving up- and downhill and the optimal motorcycle line in a curve depends on the curves before and after the current curve (Spiegel 2002). Therefore, the sensors data should be reasonably different between the driving directions. It might have been preferential to select a route where no overlaps occur. However, the battery of the notebook used for data storage restricts the maximum route length. Missing experience with battery lifetime under the test conditions led to a limitation to routes that can be finished within 1 h. The first route shown in Fig. 2 starts at the bottom right and has a length of 63 km. The first section to point 1 is through an urban area, the second section between points 1 and 2 is a rural road with only two sections in villages. Section 3 connecting points 2 and 3 is a rural road through 7 small villages. The final section from point 3 to the start/end point is again in an urban area. The second route shown in Fig. 3 starts at the top right and has a length of 67 km. The first section to point 1 starts in an urban area and continues through connected villages, the second section between points 1 and 2 is a rural road with only one section in a village. Section 3 between points 2 and 3 is a rural road with 2 small villages. The final section from point 3 to the start/end point is again in an urban area.

5 Analysis of the Test Data

The raw data is illustrated in Fig. 4. All four graphs a–d show a 100 s long fragment of the data. The upper two graphs a and b contain acceleration data. Graph a is a representation of data collected in an urban environment in a legal speed range below 50 km/h, graph b of data collected in a rural area with higher speed. Two time series are centred around the zero line, one is shifted upwards by approximately 10 m/s^2 . This last one is the acceleration along the z-axis. The other two curves represent longitudinal and lateral acceleration. Both, in graphs a and b, the lateral acceleration is the upper line. A first look at the raw data provides some insights: There is significant noise in the acceleration data due to the vibrations of the engine. Graph b has an even higher noise level than graph a due to the higher speed and higher engine rotations. There seems to be an offset between the accelerations along x- and

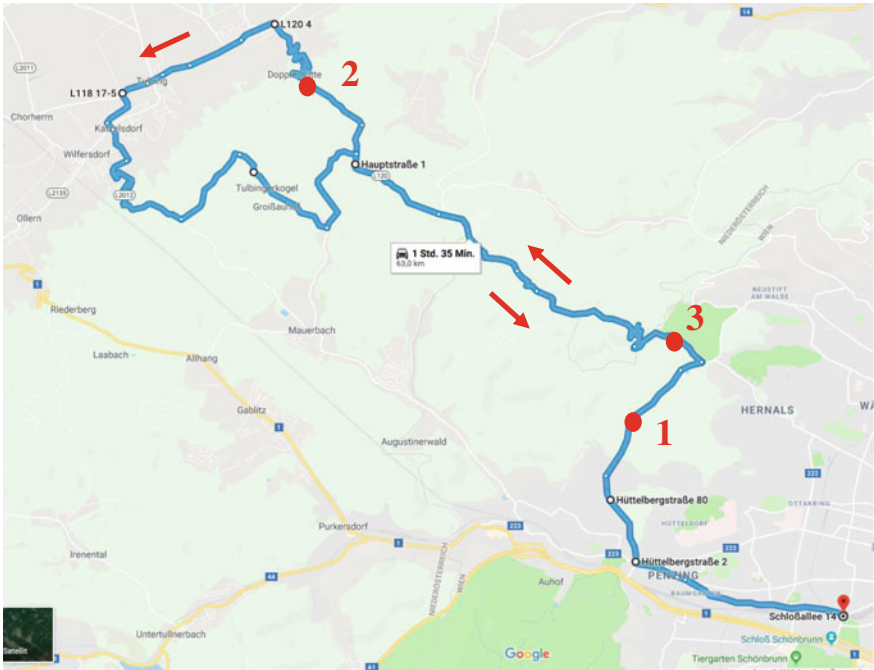


Fig. 2 Track for day 1 (Sept. 21st, 2018), track 1: <https://goo.gl/maps/5UgnzcrK8n22> (background data: © Google Maps)

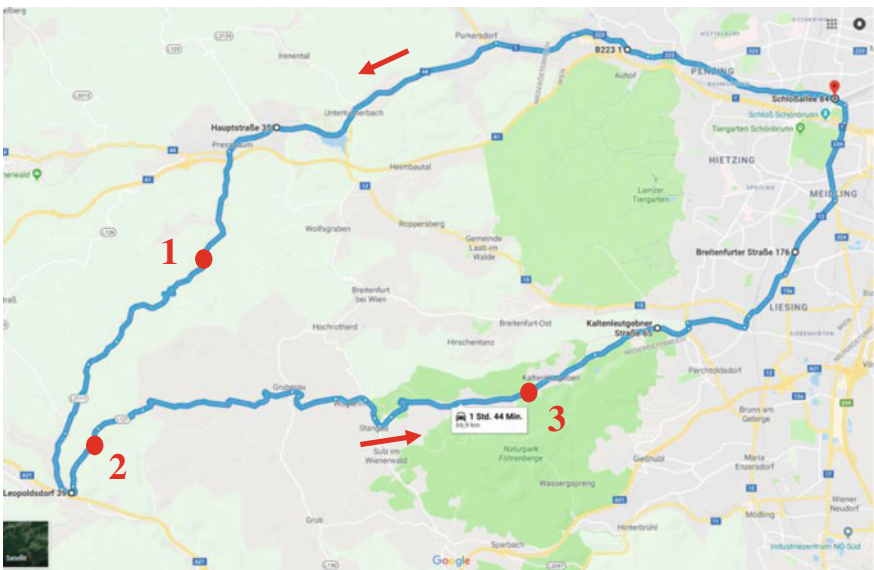


Fig. 3 Track for day 2 (Sept. 22nd, 2018), track 2: <https://goo.gl/maps/Rgjt22owKg82> (background data: © Google Maps)

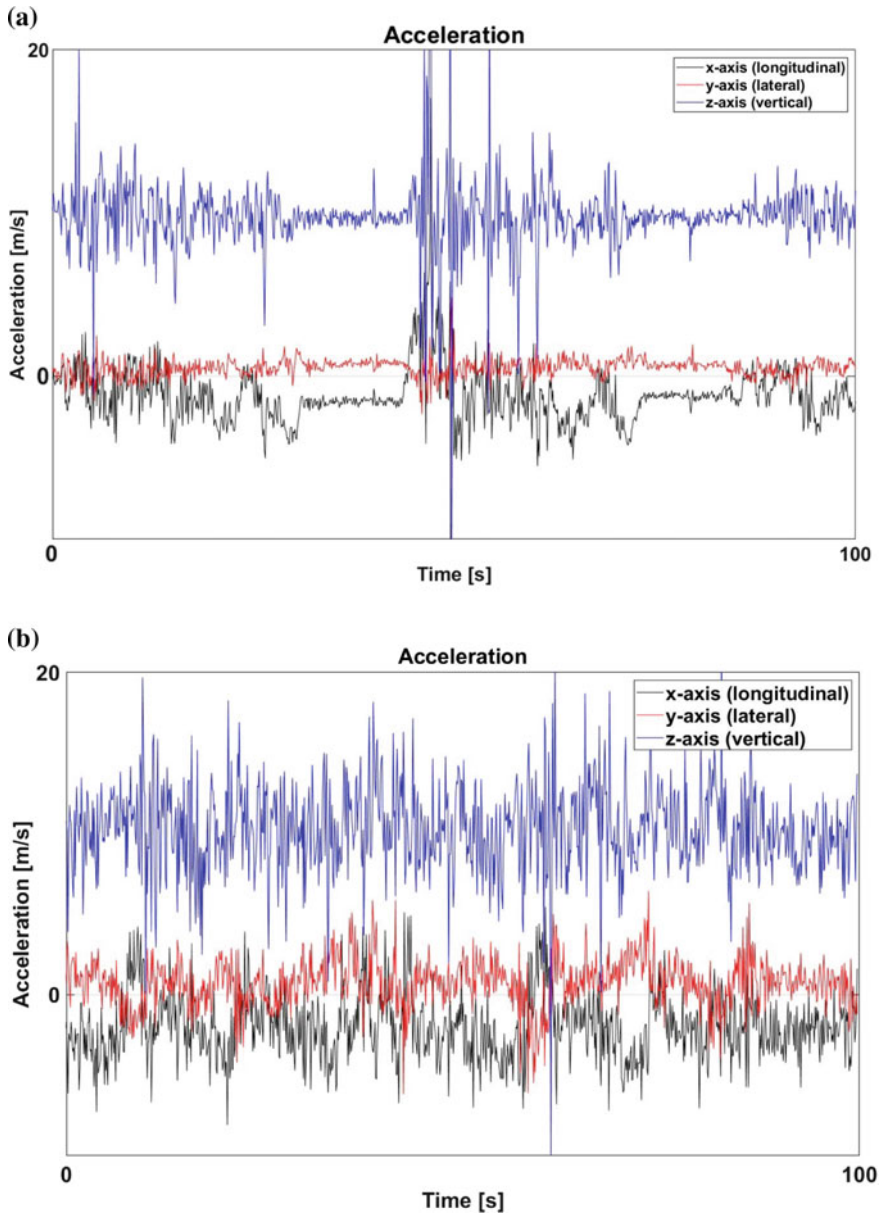


Fig. 4 Raw data series from a test drive: **a** acceleration in an urban area, **b** acceleration in a rural area, **c** rotation in an urban area, and **d** rotation in a rural area

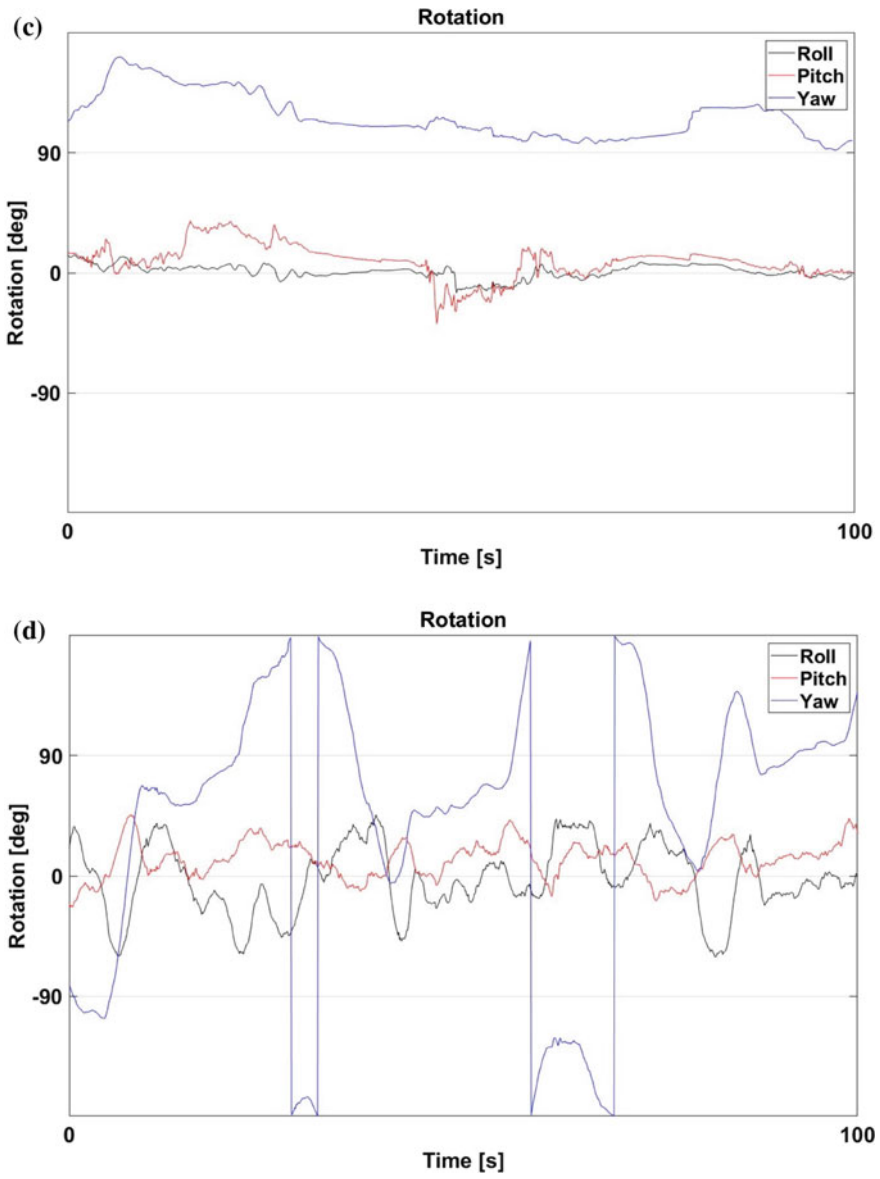


Fig. 4 (continued)

Table 1 Statistical parameters for lean angle and acceleration

	Rural/urban	Lean angle standard deviation (°)	Acceleration standard deviation (m/s ²)
1a	Urban	12.1	1.6
1b	Rural	20.1	1.9
1c	Rural	21.6	1.9
1d	Urban	14.0	1.4
2a	Urban	9.0	1.5
2b	Rural	16.5	1.9
2c	Rural	19.6	2.2
2d	Urban	9.4	1.6

y-axis. This is due to the fact that the used sensor cannot be corrected by external signal sources and the mounting on the motorcycle was not optimal compared to the integrated sensors. The other two graphs (c and d) show angular data (yaw/orientation, pitch and roll/lean angle). It is much smoother than the acceleration data.

Visual inspection of the data shows some interesting patterns:

- The acceleration graph clearly identifies waiting times, e.g., at red traffic lights. Graph a in Fig. 4 shows an example in the left half of the image.
 - The acceleration along all three axis is constant but affected by noise causing a variation in the range of approximately 0.5 m/s².
 - At the same time where the acceleration indicates a waiting time, the angular measurements (graph c) show a slight drift. Whereas a variation in lean angle and pitch are possible during waiting times at a red traffic light as a result of the rider shifting his position, the change in yaw is definitely a result of sensor drift.
 - At the end of the waiting period, a spike in pitch shows the effect of acceleration. The spike is negative due to the mounting position of the sensor.
- Roll behaves different in urban and rural areas. In urban areas, there are only occasional spikes when turning at an intersection. In rural areas, the higher driving speed required larger roll to follow the road through a curve.

The comparison between graphs c and d in Fig. 4 suggests, that roll should have different statistical parameters in rural areas than in urban areas. The total deviations from a zero angle are smaller in urban areas, which should be reflected in the standard deviation. Table 1 presents the standard deviation for the eight segments of the two routes. The standard deviation of the lean angle is between 9.0° and 14.0° in urban areas and between 16.5° and 21.6° in rural areas.

A more detailed look at the data reveals that the behaviour is not stable. Section 1c is the largest standard deviation. When cutting the whole section into pieces of 1 min driving, the standard deviations vary significantly:



Fig. 5 Twisting roads—left in section 2 of route 1, right in section 2 of route 2

17.3°	24.6°	25.9°	17.7°
10.4°	16.1°	23.7°	10.7°
17.5°	20.7°	23.5°	29.6°
27.8°	19.7°	19.1°	21.4°
14.7°	14.6°	18.9°	30.2°

Two of the values could also be in urban areas. A reason for this could be that there are several small villages along the section. Two of them apparently were long enough to show behaviour of urban areas. A separation between rural and urban areas would require one of the following pieces of information:

- driving speed (assuming that it reflects changes in the speed limit)
- absolute position (for map matching)
- map position (from a navigational device)

A remarkable difference in lean angles can be detected in two seemingly similar situations. On both tracks there are segments with twisting roads on an inclining road. The radius of the turns is comparable. This would suggest that the lean angles (shown in Fig. 6) are also comparable. On average, lean angles were around 25°. However, in one corner (marked with a circle in Fig. 5) the lean angle was only 10° (marked with a circle in Fig. 6). The reason for this difference is that there were large amounts of leaves on the tarmac. There was some rainfall during the night between the two days. Leaves in forest areas take a long time to dry so there was a high chance that the leaves were slippery and this resulted in a lower lean angle. This is an example that the data could reflect road surface properties. However, knowledge on the actual situation is necessary to interpret the data properly.

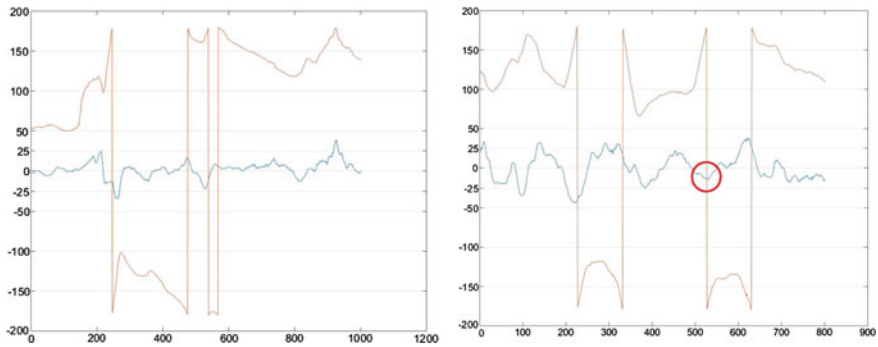


Fig. 6 Lean angles in twisting roads

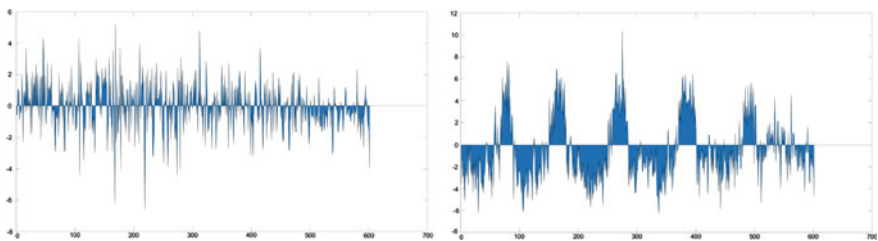


Fig. 7 Acceleration data in an urban area (left) and a rural area (right)

Another observation is the duration of acceleration and deceleration manoeuvres. In urban areas they should be much shorter and more frequent and this can also be seen in the data (compare Fig. 7). The example for the rural area was taken at a twisting road and therefore each corner requires braking before the corner and acceleration after the corner.

The areas between the two graphs seem to be different. The normalized area (divided by the number of data points) is 1.2 for the urban segment and 2.5 for the rural area. However, these are extreme cases. Derived for longer segments the difference is not that large. The four urban segments result in 1.1, 1.0, 1.1, and 1.1 whereas the rural segments result in 1.5, 1.5, 1.5, and 1.7 so there seems to be an increase of approximately 30% when moving from urban to rural segments.

6 Discussion and Further Research Questions

In this work, we present first insights demonstrating that IMU data from a motorcycle can provide information on different aspects of the road. Although the measurement of the rider's fun itself is not possible, the assumption that rider's fun relates to the amount of lean angle is reasonable. It became apparent that, in general, lean angles

have a larger variation in rural areas than in urban areas where roads tend to be straighter, traffic tends to be denser, and speed limits tend to be lower than in rural areas.

Due to the insufficient stability of the sensor, the data are imperfect. However, it was still possible to show the different characteristics of sensory data in different situations. Thus, the general feasibility of the approach was shown. Further experiments with better sensory fixation (or even usage of the internal sensors) will be necessary to improve the capacity to assess the situation from the sensory readings. It became evident that some aspects need further investigation, e.g.,

- better mounting of sensor or use of already existing, internal sensors,
- connecting the sensor to GNSS would provide absolute position and would allow map matching to identify road segment or match to data from navigation device (via time),
- better classification of situation by using a camera to identify the cause of specific behaviour (similar to Zeile et al. 2016),
- use of machine learning to analyse data, and
- use of different riders and different motorcycles to develop a strategy for normalization of the data to merge the data from different sources.

The approach can also be discussed in the context of Volunteered Geographic Information (VGI, Goodchild 2007). Each motorcycle rider can collect this kind of data. After analysis and classification the data could be stored and used together with the results of other riders to assess the suitability of a specific road segment. The collection could be either done by the manufacturers themselves but that would require information on the location of data collection. Without GNSS sensors this is a challenge due to the sensory drift. However, the data could also be analysed and communicated by a navigational service. A navigational service must have access to the location of the motorcycle and can perfectly combine sensory readings and road segments. The data could either be uploaded on the fly or in arbitrary intervals (e.g., each time the road network data are updated). However, the data should be reasonable anonymized to comply with privacy (for the problem of anonymization compare McKenzie et al. 2016). Finally, the approach is not restricted to motorcycles. Bad weather conditions also influence car drivers and collected sensory data can improve their driving security. This would be an alternative to the prediction of the situation (Litzinger et al. 2012).

References

- De Filippi P, Tanelli M, Corno M, Savaresi SM (2011) Enhancing active safety of two-wheeled vehicles via electronic stability control. In: Proceedings of the 18th world congress, The International Federation of Automatic Control, Milano, Italy, pp 638–643
- Goodchild MF (2007) Citizens as sensors: the world of volunteered geography. *GeoJournal* 69(4):211–221

- Hochmair H, Navratil G (2008) Computation of scenic routes in street networks. In: Car A, Griesebner G, Strobl J (eds) *Geospatial crossroads @ GI_Forum*. Wichmann, Berlin, pp 124–133
- Klingspor V (2018) Berechnung populärer Routen. In: Strobl J, Zigel B, Griesebner G, Blaschke T (eds) *AGIT - Journal für Angewandte Geoinformatik*. Wichmann, Berlin, pp 189–198
- Litzinger P, Navratil G, Sivertun Á, Knorr D (2012) Using weather information to improve route planning. In: Gensel J, Josselin D, Vandenbroucke D (Hrg) *Bridging the geographic information sciences*. Springer, Lecture notes in geoinformation and cartography/7418 (2012), ISBN: 978-3-642-29062-6; pp 199–214
- McKenzie G, Janowicz K, Seidl D (2016) Geo-privacy beyond coordinates. In: *Proceedings of the 2016 AGILE conference*, pp 157–175
- Navratil G (2012) Curvyness as a parameter for route determination. In: Jekel T, Car A, Strobl J, Griesebner G (eds) *GI_Forum 2012 geovizualisation, society and learning*. Wichmann, Berlin, pp 355–364
- Partusch A, Navratil G, Fiby H (2014) A Framework to determine the spatial variation of the optimal path. In: Vogler R, Car A, Strobl J, Griesebner G (eds) *GI_Forum 2014 geospatial innovation for society*. Wichmann, Berlin, pp 155–164
- Seiniger P, Winner H, Gail J (2008) Future vehicle stability control systems for motorcycles with focus on accident prevention. In: *Proceedings of the 9th Biennial ASME conference on engineering systems design and analysis*, ACSME, Haifa, Israel, 10 p
- Shannon CE (1949) Communication in the presence of noise. *Proc Inst Radio Eng* 37(1):10–21
- Spiegel B (2002) *Die obere Hälfte des Motorrads*. Motorbuch Verlag, 296 p
- Zeile P, Resch B, Loidl M, Petutschnig A, Dörrzapf L (2016) Urban emotions and cycling experience—enriching traffic planning for cyclists with human sensor data. In: Car A, Griesebner G (eds) *GI_Forum*. Wichmann, Berlin, pp 204–216

Route Choice Decisions of E-bike Users: Analysis of GPS Tracking Data in the Netherlands



Gamze Dane, Tao Feng, Floor Luub and Theo Arentze

Abstract Over the past years, the usage of electric bikes has emerged. E-bikes are suitable for short and medium distance trips. Therefore, the Dutch government promotes using e-bikes for daily commuting trips. However, the impact of increasing demand on the cycling infrastructure is unclear. Additionally, route choice models for e-bikes are limited. This paper estimates a route choice model for e-bike users in the Noord-Brabant region of The Netherlands. The data used are based on 17626 trips from 742 users including user profiles extracted from GPS data. In order to analyze the data, a mixed logit model is applied on the route choice of respondents with addition of the path-size attribute. Mixed logit model allows a panel data setup and enables the examination of preference heterogeneity around the mean of distance attribute. Moreover, the path-size attribute is included on the model to account for the overlap between alternatives. Socio-demographic characteristics and trip-related factors are found to be influencing on the route choice decisions of e-bike and bike users. There are differences on the significance of variables between e-bike and bike users.

Keywords Big data · Route choice · E-bike · GPS

G. Dane (✉) · T. Feng · F. Luub · T. Arentze
Department of Built Environment, Eindhoven University of Technology, Eindhoven,
The Netherlands
e-mail: g.z.dane@tue.nl

T. Feng
e-mail: t.feng@tue.nl

F. Luub
e-mail: f.luub@tue.nl

T. Arentze
e-mail: t.a.arentze@tue.nl

© Springer Nature Switzerland AG 2020
P. Kyriakidis et al. (eds.), *Geospatial Technologies for Local and Regional Development*, Lecture Notes in Geoinformation and Cartography,
https://doi.org/10.1007/978-3-030-14745-7_7

1 Introduction

In 2016, the motorized vehicle travel kilometers in The Netherlands has increased with 3.1% compared to the year before (Rijkswaterstaat 2017), which in turn leads to an increase in traffic congestion and average travel time. This increase has negative consequences both on the environment and the accessibility of urban regions due to air pollution and congestion. Cycling is considered to be a sustainable alternative mode to motorized vehicles, especially for short trips. Since, besides its positive impact on a person's physical and mental health, when cycling is more adopted by people, it can reduce the emission and congestion (de Hartog et al. 2010). Therefore, governments are focusing on projects to stimulate bicycle usage for commuting trips. In such projects, participants collect points per cycling kilometer or they receive a financial incentive for using bikes (Tertoolen et al. 2016).

Over the past years, the usage of electric bikes has emerged. The market share of e-bikes has increased up to almost a third of the new bike sales in the Netherlands (BOVAG-RAI vereniging 2016). This type of bicycles are equipped with a small electric motor that assists while paddling which makes it easier to travel longer distances. It also lowers the effort of cyclists to overcome natural obstacles such as elevations and wind (Wachotsch et al. 2014). Thus, it increases the average trip radius with increased speed while the rider feels less tired. These specifications make e-bike a suitable mode of transportation for medium distance trips that otherwise are made by car or public transport. However, this increase in usage of e-bikes will also increase the usage of cycling network and thus can result in the need for a better infrastructure for cycling routes. Due to that, it is important to understand the route choice decisions of e-bike users.

Route choice models are used to gain insights on travelers' preferences on route characteristics and predict traveler's behavior in the transport network (Prato 2009). In recent years, many studies have been conducted in route choice behavior of cyclists such as Landis et al. (1997), Frejinger (2008), Sener et al. (2009), Menghini et al. (2010), Broach et al. (2012), Allemann and Raubal (2015) and Ton et al. (2017). The overview of the researches conducted on the factors influencing the cycling route choice can be found in Table 1. The factors that influence route choice can be categorized in four groups: the characteristics of travelers, the route, the trip and other circumstances such as weather, safety (Bovy and Stern 1990). In order to incorporate these factors in route choice studies, either stated preference or revealed preference data are collected. Although stated preference data collection is easier and less expensive than revealed data collection, it is difficult for respondents to visualize the given choices in a real-world context (Broach et al. 2012). Revealed data collection commonly consists of travel diaries and drawing the chosen routes on maps (Altman-Hall 1996). Such data collection is quite demanding from the respondents. Moreover, the respondents may not remember the trip and the chosen route after the trip is made. This may create bias in the data or result in missing information.

Table 1 Overview factors influencing route choice behavior of cyclists

	Factor	References
Trip	Trip purpose	Altman-Hall (1996), Ben-Akiva and Bierlaire (1999), Stinson and Bhat (2003)
	Time of day and/or daylight and/or peak hours	Dill and Gliebe (2008), Li et al. (2017), Ton et al. (2017), Winters et al. (2011)
Route road	Cycling facility	Altman-Hall (1996), Bradley and Bovy (1984), Casello and Usyukov (2014), Hood et al. (2011), Hunt and Abraham (2007), Landis et al. (1997), Li et al. (2017), Menghini et al. (2010), Misra and Watkins (2017), Sener et al. (2009), Shafizadeh and Niemeier (1997), Stinson and Bhat (2003)
	Travel time and/or distance	Allemann and Raubal (2015), Bradley and Bovy (1984), Broach et al. (2012), Hunt and Abraham (2007), Li et al. (2017), Menghini et al. (2010), Misra and Watkins (2017), Sener et al. (2009), Ton et al. (2017), Usyukov (2013))
Route traffic	Obstacles (number of left turns, Stop signs and/or traffic lights)	Allemann and Raubal (2015), Ben-Akiva and Bierlaire (1999), Bierlaire et al. (2013), Broach et al. (2012), Menghini et al. (2010), Sener et al. (2009), Stinson and Bhat (2003), Ton et al. (2017)
	Volume	Bradley and Bovy (1984), Hunt and Abraham (2007), Landis et al. (1997), Li et al. (2017)
	Safety (perceived and/or actual)	Broach et al. (2012), Buehler and Pucher (2012), Casello and Usyukov (2014), Huisman and Hengeveld (2014), Hunt and Abraham (2007)
	Street lights	Misra and Watkins (2017)
Route environment	Slope (uphill and/or downhill)	Broach et al. (2012), Casello and Usyukov (2014), Dill and Gliebe (2008), Hood et al. (2011), Hunt and Abraham (2007), Landis et al. (1997), Menghini et al. (2010), Sener et al. (2009), Stinson and Bhat (2003), Zimmermann et al. (2017)
	Scenery	Landis et al. (1997), Milakis and Athanasopoulos (2014), van Overdijk (2016), Winters et al. (2011)
Traveler	Gender	Altman-Hall (1996), Dill and Gliebe (2008), Heinen et al. (2013), Stinson and Bhat (2003)
	Age	Altman-Hall (1996), Hunt and Abraham (2007), Stinson and Bhat (2003)
Other circumstances	Weather	Dill and Gliebe (2008), Romanillos et al. (2016), Stinson and Bhat (2003), Weber et al. (2014)

With recent developments in location-based services, researchers have started to use Global Positioning Systems (GPS) technologies to collect revealed preference data for route choice modeling. GPS tracking allows collecting large datasets with low cost while revealing the actual route chosen with high accuracy. Not only has the growing volume of GPS usage made it an attractive data source for researchers but also the high level of detail that it provides is influential. The current state of the technology makes it possible to collect data at an individual level every few seconds and with an accuracy of a couple of meters. This results in big datasets with many data points in time and space for each individual using GPS. Although it creates many opportunities, using the GPS data in route choice models has some difficulties: the computation time increases with big data, datasets lack of sociodemographic information, sample bias problems occur due to voluntary response, random variations and noise may occur during data collection, and there are differences in data processing algorithms. These difficulties are acknowledged by several studies such as Menghini et al. (2010), Hood et al. (2011), Usyukov (2013), Ton et al. (2017) which have used GPS data specifically for bicycle route choice modeling. Due to limitations of GPS data, they either used small sample of selected cyclists or did not include user characteristics in the route choice models.

Although, there are studies on route choice modeling by using GPS data and advanced modeling techniques, these studies do not exactly focus on e-bike users. The existing research on route choice for e-bike users is scarce and there is a lack of knowledge whether this group differs in route choice decisions from traditional cyclists. Therefore, the impact of increasing number of e-bike users on infrastructure is unclear. In order to promote the e-bike usage for short and medium distance trips through a well-designed network, it is necessary to identify which characteristics have an effect on the route choice of e-bike users. The existing studies that use GPS data on route choice of e-bike users generally have small data sets such as Allemann and Raubal (2015) or focus only on the influence of route and trip characteristics while excluding the characteristics of users such as Ton et al. (2017) because GPS data in general lacks user profiles such as gender and age of users, and only give information about start/end time of trip, latitude and longitude, speed and altitude. Thus, the objective of this research is to add to the understanding of the route choice decisions of e-bike users by presenting a route choice model estimation that includes user and trip characteristics by means of a large set of GPS data collected from the Noord-Brabant region of The Netherlands.

The remainder of the paper is organized as follows. First, the methodology is represented by describing the dataset and its processing. Then the results of the analysis are described. Finally, we conclude the paper with discussions, recommendations and possible future directions.

2 Methodology

2.1 Data Collection

The data is gathered through a dedicated mobile phone app that tracks volunteered bike and e-bike users for every day of a year since 2013. This app is designed for a project in collaboration with the Ministry of Infrastructure and Environment and the Province of Noord-Brabant in order to stimulate people to use bike or e-bike for their commuting trips by a rewarding system. Respondents had to register on the project website and download the smartphone app in order to participate in the reward program, they also had to be 18 years or older, have a minimum commute of 4 km to a destination in the Noord Brabant region and make their commute at least 50% of the time by car during the last 3 months. Currently, there are approximately 5000 people who use the app.

This data includes GPS tracks and basic user profiles such as age and gender. For this study, we use the data from 742 respondents in March, 2014, containing a total of 17626 bike and/or e-bike trips with an origin and/or destination in the Noord Brabant region. The GPS traces, obtained from the app, were transformed into activity-travel diaries by the Trace Annotator software developed by the DDSS research group of Eindhoven University of Technology. The Trace Annotator is used to recognize the transportation mode and segment the GPS traces into trips. An advantage of the Trace Annotator is the included imputation model, which is trained in the recognition of the transport mode by combining accelerometer data and GPS data, and predicts the bicycle mode with an accuracy of 97% and the e-bike mode with 99%. The GPS points were connected sequentially and matched to transportation network data acquired from OpenStreetMap (OSM). For this, an algorithm was used which first looks at the possible road segments around a GPS point and then identifies the most probable one. After the transportation mode was determined and the chosen route was matched to the transportation network, alternatives were generated and added to the dataset. Both the map matching of the GPS coordinates and the generation of alternatives based on a routing module was done prior to this study. Detailed information on the process of mode identification and map matching can be found in Feng and Timmermans (2013a, b).

The following attributes are selected to be included in the route choice model: gender, age, distance traveled, transport mode (bike or e-bike), day of the week (weekend/weekday), daylight (light/dark), peak time (peak-off/peak), route to work (PCWork) and route to home (PCHome). The variables 'age' and 'gender' are extracted directly from the project dataset. 'Weekday/weekend', 'peak-off/peak' and 'daylight' are calculated based on the start and end time of each trip found in the project dataset, compared to the peak times defined by Dutch road network (Rijkswaterstaat 2017) and the times of sunrise and sunset obtained from the database of the Dutch Meteorology Institute (KNMI 2014). There are also two attributes that represent trip purposes. 'PC work' and 'PC home' indicate whether the endpoint of the trip corresponds with the given work or home location in the project dataset.

Further indication of trip purpose is not possible due to the method of data collection and privacy reasons. For this study, bike and pedestrian paths are selected from OSM data and the paths for motorized vehicles are excluded from the dataset. Characteristics of the route in the choice model is represented only by 'distance' attribute that is extracted from OSM for both chosen route and choice alternatives. In addition, travel duration is derived also from OSM data and used for the shortest path calculation.

2.2 Choice Set Generation

In order to estimate a route choice model with revealed preference data, it is necessary to know the chosen route but also the alternative routes that are not chosen. The set of all possible routes between a given origin and a destination is called the choice set (Bovy and Stern 1990). A large group of choice alternative generation methods are based on the shortest path algorithm proposed by Dijkstra (1959). The most straight forward approach that is based on Dijkstra's shortest path algorithm is the K-shortest path algorithm. This algorithm calculates the shortest path, then the second shortest path until the desired number, K, of shortest paths is reached. The shortest path approach has the problem that it assumes perfect knowledge of the network and the shortest path between origin and destination, therefore this can be problematic for large and complex networks. However, due to its straightforward approach, we applied it in this research.

To produce the feasible choice sets, possible alternatives are generated for each observed route. The alternative routes were generated for each O-D pair through the routing component with a set of heuristic rules. The K-shortest path method was used to generate up to 5 alternatives for each trip and added to the dataset, along with the actual chosen route. In some cases, the algorithm didn't provide more than 1, 2, 3 or 4 alternatives which resulted in unequal number of alternatives for choice sets. In addition to that, the algorithm produced routes that were also the actual choices, adding the alternatives along with the actual choices to the dataset resulted in having some of the routes as both an actual choice and as an alternative. In such cases the generated alternative was removed from the choice set. Finally, all data is converted into a table format which can be used for modeling and estimation. All discrete variables are effect coded for the estimation.

2.3 Estimation Method

There is a wide variety of models used in route choice research. Most models in travel behavior studies are based on the random utility theory, which assumes that travelers try to maximize the utility and find the optimal combination of attribute

values according to their preferences, when choosing amongst alternatives. In choice set C_n of individual n the utility U_{in} of alternative i is given by:

$$U_{in} = V_{in} + \varepsilon_{in} \quad (1)$$

the deterministic term V_{in} consists of individual characteristics and alternative attributes. The random error term ε_{in} is incorporated to account for uncertainty caused by unobserved individual characteristics, unobserved attributes or measurement errors. To estimate the probability that a certain alternative is chosen, several different models have been proposed in literature for route choice modeling. They can be grouped by the way they deal with the overlapping problem that is created when alternative routes share a link or multiple links. The first and most basic group of models are those that do not account for overlap amongst alternatives, for example the Multinomial Logit or Nested Logit. Second is the group of models that use a tree structure and account for overlap through the random error component of the utility function, examples are Cross Nested Logit and the Generalized Nested Logit. These models overcome the problem of correlation amongst nests, but they both create an extremely large and complex model structure when applied to real networks. The third group of models allow accounting for the overlap amongst alternatives by adding an extra attribute to the deterministic part of the utility function of the model. An example of this is the Path-Size Logit which is very often used in route choice modelling due to its easiness to calculate and low computational effort. The attribute accounting for the overlap amongst the alternatives is called the path-size (PS) attribute as shown in Eq. 2.

$$U_{in} = V_{in} + \beta_{PS} \ln PS_{in} + \varepsilon_{in} \quad (2)$$

With

$$PS_{in} = \sum_{a \in \Gamma_i} \frac{L_a}{L_i} \frac{1}{\sum_{j \in C_n} \delta_{aj}} \quad (3)$$

where Γ_i is the set of links in path i ; L_a and L_i are the length of link a and path i . δ_{aj} is the link-path incidence variable that is 1 if link a is on path j and 0 otherwise.

The fourth and last group contains the models that account for overlap by allowing covariance between the error terms of the alternatives, such as Mixed Logit model (ML). The advantage of this model over the other models is that it allows alternatives to be correlated, it allows coefficients to be randomly distributed which enables understanding heterogeneity across respondents and it is computationally feasible for route choice decisions.

It is important to note here the data structure that is used for the estimation in this research. Due to the alternative generation method, the alternatives in the choice set are unlabeled. So that the alternatives are described by the values of their attributes. Only the “distance” attribute varies over the alternatives and the rest of the attributes

are fixed. Due to that, all fixed attributes are entered in the model as interaction with “distance”. Moreover, as each respondent can have more than one trip registered, the data is treated as panel data. In this case the panel structure is unbalanced due to different numbers of entries per respondent. In short, the dataset consists of unlabeled variable choice set with unbalanced panels. To account for the panel structure, an ML model is applied in this research with addition of the path-size attribute because the path-size attribute is necessary to account for the overlap between alternatives. Path-size attribute is calculated as in Eq. 3. Using this model we estimate a distribution for the distance parameter to examine the preference heterogeneity around the mean of distance attribute. In this paper, we estimate the ML model with 1000 Halton draws on the samples of bike trips and e-bike trips separately by using Nlogit software.

3 Results

3.1 Descriptive Analysis

Table 2 shows the distribution of age and gender across the sample. In the dataset, females are slightly more represented than males. Moreover, respondents above 44 years old are highly represented in the dataset. This might be due to the rule of the project that participants have to be over the age of 18 and the project focused on commuters, this causes the age category of younger than 35 years to be very small compared to the overall sample. In addition, the oldest respondents in the sample is recorded to be 66 years old. Table 3 shows the trip frequencies per traveler characteristic and mode. We can say that more e-bike trips has been conducted by respondents. Some respondents conducted both bike and e-bike trips at different times. Females have conducted more bike and e-bike trips than males. The majority of the trips are conducted by respondents between 45 and 66 years old.

Table 4 shows the descriptive statistics of the distance and its natural log transformations per transport mode. On average, respondents traveled 5.5 km by bike and 9 km by e-bike. Table 4 also describes the path-size variable. Since a path-size of 1

Table 2 Distribution of socio-demographic variables across sample

Gender	Sample	Percentages (%)
Male	354	48
Female	388	52
<35 years	33	4
35–44 years	130	18
45–54 years	346	47
>55 years	233	31
Total	742	100

Table 3 Trip frequencies per traveler characteristic and mode

Mode	Age	Male	(%)	Female	(%)	Total	(%)
Bike	<35 years	23	2	56	4	79	3
	35–44 years	175	16	338	22	513	19
	45–54 years	431	39	710	46	1141	43
	>55 years	485	44	431	28	916	35
	Total	1114	100	1535	100	2649	100
E-bike	<35 years	135	2	352	4	487	3
	35–44 years	1198	17	1215	15	2413	16
	45–54 years	3045	43	4178	53	7223	48
	>55 years	2760	39	2094	27	4854	32
	Total	7138	100	7839	100	14977	100
Total	<35 years	158	2	408	4	566	3
	35–44 years	1373	17	1553	17	2926	17
	45–54 years	3476	42	4888	52	8364	47
	>55 years	3245	39	2525	27	5770	33
	Total	8252	100	9374	100	17626	100

Table 4 Trip frequencies per route characteristic and mode

	Minimum	Maximum	Mean	Std. deviation
<i>Bike</i>				
Distance (km)	0.08	29.05	5.46	5.04
ln (distance)	-2.50	3.37	1.19	1.10
Path Size	0.28	1.00	0.84	0.14
ln (Path Size)	-1.29	0.00	-0.19	0.19
<i>E-bike</i>				
Distance (km)	0.12	31.34	9.04	5.51
ln (distance)	-2.15	3.44	1.93	0.84
Path-Size	0.22	1.00	0.82	0.16
ln (Path-Size)	-1.50	0.00	-0.23	0.22

means a unique route and of 0 complete overlap, a mean of 0.84 (0.82) and standard deviation of 0.14 (0.16) indicates that the alternative generation algorithm performed rather well in creating different alternatives with only some overlap.

3.2 Estimation Results

The sample is divided into two: bike and e-bike trips, and we estimated the ML model on each sample including the path-size attribute. The distance parameter is included as a random parameter in the model. As each attribute, except path-size, is interacted with the distance attribute, the results reflect the route choice of respondents based on the distance attribute. According to the results as shown in Table 5, pseudo rho squared values are 0.340 for the bike model and 0.482 for the e-bike model. Pseudo rho squared values greater than 0.3, which is the case for both models, are considered good (Hensher et al. 2005).

For both bike and e-bike models, the distance and the path-size are found to be significant. 'Distance' is the only variable in the model which has a main effect since all independent variables are fixed over the alternatives in the choice set. Its effect is significant in all models however its sign is counterintuitive. The positive sign of the estimated coefficient means that if the distance of a route increases so does the probability of the route being chosen, suggesting that people prefer longer routes. This is not in line with findings in existing literature. The estimate for the standard deviation of distance is found to be positive and significant for both bike and e-bike, meaning that for both bike trips and e-bike trips taste variation for distance exists in the sample.

The path-size variable estimate is found to be positive and significant at a 1%-level, as expected. The sign needs to be positive in order to correct the utility for the overlap amongst alternatives and the fact that the estimates have a significant effect proves that there is indeed correlation between the alternatives which have overlap.

Besides distance and path-size, the age category '35–44 years' and the 'daylight' estimate are found to be significant in the model for bike trips. The age category estimate has a negative sign and is significant at 10% level. This means that respondents between 35 and 44 years old prefer a longer distance to a lesser extent than respondents over 55 years old, which is the base category. The 'daylight' estimate is highly significant, at 1% level, and has a positive sign. This means that for bike trips, the likelihood of a longer route being chosen at night is smaller than during the day.

For e-bike trip model, besides distance and path-size, the estimates for age category '45–55 years', 'weekday', 'peak' and 'PCwork' are found to be significant. Age category '45–55 years' is significant at 5% level while others are significant at 1% level. The negative estimate for the age category '45–55 years' indicates that individuals in this age category prefer a longer distance to a lesser extent during e-bike trips than individuals over 55 years old, which is the base category. The negative estimate for 'weekday' indicates that during weekdays respondents prefer a longer distance to a lesser extent than during the weekend when using an e-bike. The negative estimate of 'peak' indicates that when respondents are making a trip using an e-bike during peak hours they prefer a longer distance to a lesser extent than when making an e-bike trip outside peak hours. The positive sign for the 'PCwork' estimate suggests that if the end location of the trip is the work location, respondents

Table 5 Estimation results of models for bike and e-bike

Attribute	Bike	E-bike
<i>Main effects</i>		
Ln (distance)	4.786***	4.889***
<i>Interaction effects</i>		
Gender	-0.113	-0.024
Age < 35	2.246	-1.605
Age 35–44	-1.550*	-1.731
Age 45–55	-0.826	-1.931**
Weekday	0.221	-0.831***
Peak	-0.103	-0.484***
Daylight	0.938***	-0.171
PC work	0.420	0.608***
PC home	-0.041	0.194
ln(PathSize)	10.20***	15.50***
Std. dev. of random ln(dist) coeff.	4.366***	8.780***
Log-likelihood model	-2241	-11600
Log-likelihood base	-3394	-22389
Pseudo rho-squared	0.340	0.482

Note ***, **, * Significance at 1%, 5%, 10% level

have a higher preference for longer distance routes than when the end location is not the work location.

3.3 Discussions

Comparing the models for the different transport modes shows that the significant positive estimate for distance is similar for both bike and e-bike models. This indicates that the probability of a route being chosen increases when distance increases, this contradicts the base assumption of route choice modelling, and the expectation that people prefer the shortest route. A less negative estimate for the distance coefficient might be explained by respondents’ imperfect knowledge of the network, they might not know a shorter path exists (Prashker and Bekhor 2004; Prato 2009), but this cannot explain a positive sign for the estimate. The positive estimate for distance may be caused by the alternative generation algorithm in combination with that only one route specific variable (distance) is included in the model. The generation algorithm used a k-shortest path algorithm which generated alternatives that are clustered around the shortest path, with little variety on other road attributes. The high path-size variable coefficient, which corrects the alternatives for overlap, indicates that there are unobserved variables that have a negative influence in the utility function (Broach

et al. 2012). An example of this is that many of the shortest paths generated might all pass through a high density urban area and share a busy road segment with high motorized traffic volumes and no separate bicycle lane, while the chosen longer routes detour around the high density urban area, through a more rural landscape, and have little motorized traffic or a separate bicycle path. Moreover, geometric complexity of the routes and also participants' knowledge of the available routes might influence the chosen routes. In our study, these variables are not included in the models due to lack of data while other studies found these attributes have a significant effect. The effect of such unobserved variables is captured by correlations with the distance attribute that is included in the model, which might have contributed to the positive estimate for 'distance'. An incentive created by the project might also contribute to the higher preference for longer distance routes of participants. Part of the project is that participants are rewarded for every kilometer they travel by bike or e-bike, travelling longer distances will result in a bigger reward.

Both 'peak' and 'weekday' have a significant negative effect on the utility of a route for trips made with an e-bike but not for trips made with a bike. This indicates that when respondents make a trip by e-bike they prefer a longer distance to a lesser extent during peak hours and weekdays while for respondents making a trip by bike 'peak' and 'weekday' don't have this effect. Weekday and peak hours traffic is characterized by commuters who are more time-constrained (Ton et al. 2017). The higher average speed, one of the main reasons for using an e-bike (Weinert et al. 2007), might suggest that this group of travelers is more time conscience.

The variable 'daylight' has a significant positive effect in the bike model but not in the e-bike model. This suggests that people who travel by bike indeed have a different preference for distance when it comes to whether it is day or night, but people who travel by e-bike don't have this difference in preference. The change in preference for different daylight conditions for bike trips is explained by the perceived unsafe conditions found by Stinson and Bhat (2003) and Gatersleben and Appleton (2007). This unsafety during the night is not experienced by e-bike trips.

The 'PCwork' estimate has a positive significant effect on the preferred distance for e-bike trips. The positive estimate is counterintuitive when assuming that the majority of travelers during peak hours are commuters travelling to work. However, a correlation test shows that a relation between these two variables is not that clear. The positive estimate might be explained by trip-chaining, people drop their kids off at school and travel further to work, but majority of people travelling to work do this directly (Department for Transport 2014). Why the estimate is only significant for e-bike trips and not for bike trips is likely because respondents making a trip by e-bike in general travel, and prefer, longer distances than respondents making a trip by bike. The significance of the 'PCwork' estimate indicates that when respondents are using an e-bike travelling to work, other route characteristics play a more important role in their route choice decisions.

4 Conclusions

Understanding the trends of e-bike users and their route choice behavior is important to make correct policy decisions for stimulating the usage of it. The smartphone based GPS data allows incorporating larger spatial and temporal coverage with big volumes of data. Such data can be enriched by fusing other existing datasets in terms of including relevant factors influencing route choices. That way it helps for better understanding of the behaviors and decisions of e-bike users.

This study distinguishes itself from other cyclists' route choice studies by the size of the final dataset and that there are several socio-demographic variables included that often lack in other cyclists' route choice studies using GPS data. For this study, a GPS dataset acquired from a national project and combined with network data acquired from OpenStreetMap was used. Extra variables were added with the use of secondary data on sunset-sunrise times and peak hours acquired from KNMI and Rijkswaterstaat, respectively. Moreover, this study used a Mixed-Path Size logit model, which handles the overlap problem between alternatives and takes into account the correlation caused by the repeated choice of individuals that is often ignored in other studies. Thus, the model allowed to examine whether there is any taste variation between people for the valuation of trip distance.

The estimation results show that socio-demographic characteristics and trip-related factors are influential on the route choice of both bike and e-bike users. The findings suggest that e-bike users are sensitive to distance of trips and their taste varies in terms of distance. Moreover, although usage of e-bikes is substantial within the elderly group, increasing age is negatively associated with the distance of the chosen route. In the dataset, it is also seen that respondents can be both bike and e-bike user depending on the purpose of their trips. The findings suggest that there is a difference between the preferences of bike and e-bike users' route choice behavior.

Although findings of existing studies show that multiple route characteristics are potential influencers on route choice, only distance is included in our study. In order to reduce the likely effects of unobserved variables on the route choice, multiple route characteristics should be included in the data through secondary sources such as OpenStreetMap. For instance, busy road segments or topographic complexity might also impact the route choice for bike users. So far the study has taken into account the user behavior perspective, including more spatial characteristics is necessary for future work due to the interdependency between human behavior and spatial planning. Because, topology matters in determining human activities, and in better understanding urban structure and dynamics (Ma et al. 2018). In addition, as new services are available via smartphones such as maps and navigation apps, that even provide real-time information depending on the area, might influence the route choices of bike and e-bike users. Therefore, information on whether the respondents have knowledge of the network or use navigation assistance should be incorporated to the data collection in the future for a more realistic estimation and understanding of the reality. Finally, the k-shortest path algorithm that was used to generate the alternatives may not be an adequate model of the route alternatives individuals

actually consider when making a trip. This likely influenced the estimation results of the model. Therefore, in the future, it is necessary to examine alternative route generation algorithms and test whether they lead to more realistic choice sets.

References

- Allemann D, Raubal M (2015) Usage differences between bikes and E-bikes. In: Bacao F, Santos M, Painho M (eds) *Geographic information science as an enabler of smarter cities and communities*. Springer, Heidelberg, pp 201–217
- Altman-Hall L (1996) *Commuter bicycle route choice: analysis of major determinants and safety implications*. PhD Dissertation
- Ben-Akiva M, Bierlaire M (1999) Discrete choice methods and their applications to short-term travel decisions. In: Hall R (ed) *Handbook of transportation science*. Kluwer (1999), pp 5–34
- Bierlaire M, Chen J, Newman J (2013) A probabilistic map matching method for smartphone GPS data. *Transp Res Part C Emerg Technol* 26:78–98. <https://doi.org/10.1016/j.trc.2012.08.001>
- BOVAG-RAI vereniging (2016) *Mobiliteit in Cijfers Tweewielers 2016–2017*. Stichting BOVAG-RAI Mobiliteit, Amsterdam
- Bovy P, Stern E (1990) *Route choice: wayfinding in transport networks*. Springer Netherlands
- Bradley MA, Bovy PHL (1984) A stated preference analysis of bicyclist route choice. In: PTRC annual meeting, London, pp 39–53
- Broach J, Dill J, Gliebe J (2012) Where do cyclist ride? A route choice model developed with revealed preference GPS data. *Transp Res Part A* 46:1730–1740
- Buehler R, Pucher J (2012) Cycling to work in 90 large American cities: new evidence on the role of bike paths and lanes. *Transportation* 39(2):409–432. <https://doi.org/10.1007/s11116-011-9355-8>
- Casello JM, Usyukov V (2014) Modeling cyclists' route choice based on GPS Data. *Transp Res Rec J Transp Res Board* 2430:155–161. <https://doi.org/10.3141/2430-16>
- Department for Transport (2014) *National travel survey factsheet: trip chaining*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/509447/nts-trip-chaining.pdf
- Dijkstra EW (1959) A note on two problems in connexion with graphs. *Numerische Mathematik* 269–271
- Dill J, Gliebe J (2008) Understanding and measuring bicycling behavior: a focus on travel time and route choice. *Bicycling*. <https://doi.org/OTREC-RR-08-03>
- Feng T, Timmermans HJP (2013a) Transportation mode recognition using GPS and accelerometer data. *Transp Res Part C* 37:118–130
- Feng T, Timmermans HJP (2013b) Map matching of GPS data with Bayesian belief networks. *J East Asia Soc Transp Stud* 10:100–112
- Frejinger E (2008) *Route choice analysis: data, models, algorithms and applications*
- Gatersleben B, Appleton KM (2007) Contemplating cycling to work: attitudes and perceptions in different stages of change. *Transp Res Part A Policy Pract* 41(4):302–312. <https://doi.org/10.1016/j.tra.2006.09.002>
- de Hartog J, Boogaard H, Nijland H, Hoek G (2010) Do the health benefits of cycling outweigh the risks? *Environ Health Perspect* 1109–1116
- Heinen E, Maat K, van Wee B (2013) The effect of work-related factors on the bicycle commute mode choice in the Netherlands. *Transportation* 40(1):23–43. <https://doi.org/10.1007/s11116-012-9399-4>
- Hensher D, Rose J, Greene W (2005) *Applied choice analysis*
- Hood J, Sall E, Charlton B (2011) A GPS-based bicycle route choice model for San Francisco, California. *Transp Lett* 3(1):63–75

- Huisman DJ, Hengeveld J (2014) Opmaak van een methode voor fietstellingen in de provincie Antwerpen. <https://doi.org/10.1007/s13398-014-0173-7.2>
- Hunt JD, Abraham JE (2007) Influences on bicycle use. *Transportation* 34(4):453–470. <https://doi.org/10.1007/s11116-006-9109-1>
- KNMI (2014) Tijden van zonopkomst en –ondergang 2014 52°00'. <http://www.sciamachy-validation.org/klimatologie/achtergrondinformatie/zonop2014.pdf>
- Landis B, Vattikuti V, Brannick M (1997) Real-time human perceptions: towards a bicycle level of service. *Transp Res Rec* 1578:119–126
- Li S, Muresan M, Fu L (2017) Cycling in Toronto: route choice behaviour and implications to infrastructure planning. *Transp Res* 5998
- Ma D, Omer I, Osaragi T, Sandberg M, Jiang B (2018) Why topology matters in predicting human activities. *Environ Plan B Urban Anal City Sci*
- Menghini G, Carrasco N, Schüssler N, Axhausen K (2010) Route choice of cyclists in Zurich. *Transp Res Part A Policy Pract* 754–765
- Milakis D, Athanasopoulos K (2014) What about people in cycle network planning? Applying participative multicriteria GIS analysis in the case of the Athens metropolitan cycle network. *J Transp Geogr* 35:120–129. <https://doi.org/10.1016/j.jtrangeo.2014.01.009>
- Misra A, Watkins K (2017) Modeling cyclists' willingness to deviate from shortest path using revealed preference data, vol 1, pp 1–16
- Prashker JN, Bekhor S (2004) Route choice models used in the stochastic user equilibrium problem: a review. *Transp Rev* 24(4):437–463. <https://doi.org/10.1080/0144164042000181707>
- Prato CG (2009) Route choice modeling: past, present and future research directions. *J Choice Model* 2(1):65–100
- Rijkswaterstaat (2017) Publieksrapportage Rijkswegennet, jaaroverzicht 2016. Retrieved from Rijkswaterstaat: <https://www.rijksoverheid.nl/documenten/rapporten/2017/02/14/publieksrapportage-rijkswegennet-jaaroverzicht-2016>, 30 Jan 2017
- Romanillos G, Zaltz Austwick M, Ettema D, De Kruijf J (2016) Big data and cycling. *Transp Rev* 36(1):114–133. <https://doi.org/10.1080/01441647.2015.1084067>
- Sener IN, Eluru N, Bhat C (2009) An analysis of bicycle route choice preferences in Texas, US. *Transportation* 36(5):511–539
- Shafizadeh KR, Niemeier DA (1997) Bicycle journey-to-work: travel behavior characteristics and spatial attributes. *Transp Res Board* 1578:84–90
- Stinson M, Bhat CR (2003) An analysis of commuter bicyclist route choice using a stated preference survey. *Transp Res Rec J Transp Res Board* 634(512)
- Tertoolen G, Ruijs K, de Vree R, Stelling C (2016) Change is cool. Inzichten uit fietsstimuleringsprojecten
- Ton D, Cats O, Duives D, Hoogendoorn S (2017) How do people cycle in Amsterdam? Estimating cyclists' route choice determinants using GPS data from an urban area. In: The 96th annual meeting of the transportation research board: the 2017 compendium of papers, pp 1–12
- Usuykov V (2013) Development of a cyclists' route-choice model: an Ontario case study. Dissertation
- van Overdijk RPJ (2016) The influence of comfort aspects on route- and mode-choice decisions of cyclist in the Netherlands. Thesis
- Wachotsch U, Kolodziej A, Specht B, Kohlmeyer R, Petrikowski F (2014) Electric bikes get things rolling. Federal Environment Agency (UBA), Dessau-Rosslau
- Weber T, Scaramuzza G, Schmitt KU (2014) Evaluation of e-bike accidents in Switzerland. *Accid Anal Prev* 73:47–52. <https://doi.org/10.1016/j.aap.2014.07.020>
- Weinert J, Ma C, Cherry C (2007) The transition to electric bikes in China: history and key reasons for rapid growth. *Transportation* 34(3):301–318. <https://doi.org/10.1007/s11116-007-9118-8>

- Winters M, Davidson G, Kao D, Teschke K (2011) Motivators and deterrents of bicycling: comparing influences on decisions to ride. *Transportation* 38(1):153–168. <https://doi.org/10.1007/s11116-010-9284-y>
- Zimmermann M, Mai T, Frejinger E (2017) Bike route choice modeling using GPS data without choice sets of paths. *Transp Res Part C Emerg Technol* 75:183–196. <https://doi.org/10.1016/j.trc.2016.12.009>

Route Optimisation for Winter Maintenance



Nikmal Raghestani and Carsten Keffler

Abstract In many countries, winter maintenance is a requirement to keep public life going throughout the cold season. This paper investigates the optimization of salt spreading routes in Denmark in terms of service time and cost. It looks at salting as a capacitated arc routing problem and proposes a greedy randomized adaptive search procedure to this end. At the core of the proposed approach is a heuristic algorithm based on simulated annealing that improves the initial route by searching for alternatives within a predefined search space, taking into account a number of constraints and criteria at each iteration of the procedure. The performance of the optimization approach is tested on three different existing service routes, where it is shown to reduce route length by an average of 8.7% and service time by an average of 9.5%.

Keywords Capacitated arc routing problem · Simulated annealing · Route optimization

1 Introduction

The nordic winter weather requires the authorities to remove snow and take measures against road icing in order to maintain accessibility and traffic safety. In Denmark, The Danish Road Directorate (Vejdirektoratet, VD) and the municipalities are responsible for these tasks. This paper investigates the potential for optimisation of the winter maintenance services in terms of service time and cost. We focus on salting, as this is by far the most common winter road maintenance activity (Knudsen et al. 2014). Planning of winter maintenance activities consists of various types of decision-making problems at the strategic (facility locations such as plants, depots, material

N. Raghestani · C. Keffler (✉)
Department of Planning, Aalborg University Copenhagen, Copenhagen, Denmark
e-mail: kessler@plan.aau.dk

N. Raghestani
e-mail: nikmaljuve@gmail.com

© Springer Nature Switzerland AG 2020
P. Kyriakidis et al. (eds.), *Geospatial Technologies for Local and Regional Development*, Lecture Notes in Geoinformation and Cartography,
https://doi.org/10.1007/978-3-030-14745-7_8

storages), tactical (fleet size, fleet combination and other vehicle/material related issues) and operational (determination of routing and scheduling of each vehicle) levels (Bodin et al. 1981). These levels are interconnected, but the strategic and tactical levels are often bound by a number of inflexible circumstances, so that the largest potential for optimisation is in the vehicle routing at the operational level.

This routing problem is based on a set of arcs that a fleet of vehicles has to traverse. The objective is to determine feasible routes for each vehicle with respect to the cost variables. Since one of the cost variables in road salting is the salt capacity on a vehicle, this kind of vehicle routing problem is a Capacitated Arc Routing Problem (CARP) (Eiselt et al. 1995). Since CARPs fall into the class of NP-hard, *non-deterministic polynomial-time hard* problems (Golden and Wong 1981), heuristic procedures are required in order to approach a near-optimal solution within acceptable time. This paper applies an approach based on Simulated Annealing (SA) for this purpose.

Figure 1 shows one of the winter maintenance routes served by VD. During the optimization process, the route itself, shown in green, cannot be changed due to service contracts that VD and the municipalities have with private road maintenance firms, which they can bid on in public tenders. The optimisation is therefore limited to the routing between the depot and the route, the order in which the green segments are serviced, as well as the *deadheading* during turns and between different segments of the route, during which no salt is brought onto the road. Despite these limitations



Fig. 1 Winter maintenance route K72, West of Copenhagen, with depot location in Roskilde

for route optimisation, manual inspection of the existing routing instructions has shown potential for optimisation due to the fact that the current routing instructions, including deadheading and getting from depot to route, have been created manually. The remainder of this paper will show that despite these limitations for optimisation, a heuristic approach based on simulated annealing could reduce route length by an average of 8.7% and service time by an average of 9.5% across a set of three test routes.

The next section introduces relevant related work. Section 3 formalizes the problem, followed by an introduction of the heuristic algorithm in Sect. 4. Section 5 outlines the implementation and presents the achieved results, followed by concluding remarks in Sect. 6.

2 Related Work

Scientific research on winter maintenance in terms of road salting and snow removal first began in the 1960s. In the 1980s and 1990s, winter maintenance as Arc Routing Problems (ARPs) got more attention (Lemieux and Gampagna 1984; Evans and Weant 1990; Goode and Nantung 1995; Haslam and Wright 1991; Gendreau et al. 1992). Eglese (1994) conducted research with the objective to determine the most cost-effective management of salt spreading. The approach was based on a heuristic algorithm with respect to practical constraints such as service time, deadheading time, vehicle capacities, multiple depot locations, etc. The heuristic procedure was initiated with a construction algorithm, first introduced by Male et al. (1977), and an SA heuristic was then used for improvement. The study showed that the use of SA improved the initial results (Eglese 1994).

The idea of SA was introduced by Metropolis et al. (1953) based on the cooling of molten materials, where the cooling rate determines the structural properties of the solid mass. Metropolis et al. (1953) proposed the use of SA in order to simulate the physical system's cooling process and change in energy until a state of thermal equilibrium is obtained. This is ensured with the stopping conditions of either a minimum temperature or a maximum number of iterations, thus retaining the possibility of accepting a non-improving solution in order to escape *local* minima while searching for a *global* minimum. Kirkpatrick et al. (1983) proposed the use of SA to search for near-optimal solutions in the context of routing. They proposed that the search be carried out in a wide solution space with respect to decreasing probability depending on the progresses of the algorithm by moving arcs between the routes.

The Capacitated Arc Routing Problem (CARP) has been defined based on a connected undirected graph $G = (N, A, C, Q)$, where N is a set of nodes, and A is a set of arcs. C denotes a cost vector, indicating the cost of traversing each arc, and Q is a corresponding demand matrix, indicating the demand for each arc. Given a number of identical vehicles with capacity W , the problem is to find a number of tours so that (1) Each arc with positive demand is serviced by exactly one vehicle, (2) The

sum of demand of those arcs serviced by each vehicle does not exceed W , and (3) The total cost of the tours is minimized (Golden and Wong 1981).

Existing heuristic algorithms for CARP can be divided into construction methods and improvement methods. Simple construction methods include the *Construct-Strike Algorithm* (Christofides 1973; Pearn 1989), which is one of the simplest CARP construction methods. This method progressively constructs and removes feasible routes from the original graph without disconnecting the graph until no more feasible routes can be determined. The *Path Scanning Algorithm* (Bodin et al. 1981) uses a set of five arc selection rules. This is done by selecting one rule at a time and determining the optimal routes for each of the five rules. This method was extended by Pearn (1989), who suggested random use of the rules and then selecting the optimal solution. Furthermore, Bodin et al. (1981) proposed an *Augment-Merge Algorithm* for CARP which heuristically creates feasible routes based on incomplete routes. Using the same method, Chapleau et al. (1984) proposed the *Parallel-Insert Algorithm*, which routes in parallel, while balancing the costs of the different routes. Moreover, there are a number of heuristics that follow a two-phase construction approach by either clustering route segments first and then creating routes from those clusters, or vice versa (Eiselt et al. 1995).

The heuristics for these construction methods are problem-dependent and often too greedy. This can cause trapping in a local minimum, thus failing to obtain the global minimum. To avoid this, there is a need for meta-heuristics that are not greedy and can accept a temporary solution deterioration. This allows exploration of the solution space in order to find a better solution towards the global minimum. The meta-heuristics are improvement methods for an initially constructed solution and there are a number of meta-heuristics for various problem statements within the CARP. In this paper, we use a two-phase algorithm consisting of a construction phase that generates an initial route and a local search phase, where an SA-based meta-heuristic is used to improve the initial route. While we focus on an SA-based approach in this paper, recent work on the CARP also employs genetic algorithms (Arakaki and Usberti 2018; Kurniawan et al. 2015).

3 Problem Formulation

In order to generalise the approach presented in this paper beyond salt spreading in Denmark, the development will be based on the following problem formulation:

Given a graph with both directed and undirected arcs, some of which may not require service but can be used for deadheading, and given a number of vehicles with different capacities, find a total least time set of routes that start and end at the concerned depot, while satisfying the distance, capacity, time and frequency constraints.

An existing exact approach to this kind of combinatorial optimization is integer linear programming (Nemhauser and Wolsey 1988). Integer linear programs (ILPs) can be solved with existing mathematical solvers, which allow computing glob-

Table 1 Parameters, variables and indexes used in the problem formalization

Notation	Definition	Unit
A	The network's set of arcs	[number]
C_{ij}	Distance between the nodes	[m]
D	The maximum total distance a vehicle can cover	[m]
i, j	Node index	[number]
F_{ij}^p	This decision variable ensures the elimination of sub-tours and is assigned to all arc (i, j) and vehicle p in the direction i to j	[number]
G_{ij}	The difference between deadheading time and servicing time for an arc (i, j) in the direction i to j	[minutes]
k	The number of available vehicle(s) for service	[number]
L_{ij}^p	This variable indicates if an arc needs to be serviced and $L_{ij}^p = 1$ if an arc (i, j) has to be serviced by a vehicle p in the direction i to j	[number]
n	The total number of nodes in the network	[number]
N_{ij}	The number of times an arc (i, j) in the direction i to j needs to be serviced	[number]
p	Vehicle index	[number]
Q_{ij}	The demanded salt for an arc (i, j) in the direction i to j	[m ² * $\frac{\text{kg}}{\text{m}^2}$]
T	The maximum total time a vehicle can cover	[minutes]
T_{ij}	Time for travel without servicing (deadheading), arc (i, j) in the direction i to j	[minutes]
W	Vehicle capacity	[kg]
X_{ij}^p	This variable indicates if an arc is "used" and $X_{ij}^p = 1$ if an arc (i, j) is used by vehicle p for either servicing or deadheading in the direction i to j	[number]

ally optimal solutions for small problem instances. However, since our real-world problem instances are large and the existing algorithms for solving ILPs have an exponential worst-case runtime, we developed a heuristic method based on simulated annealing. Nevertheless, in the following, we define the problem in the form of an ILP.

The problem formalization is based on a model proposed by Golden and Wong (1981) and modified by Haghani and Qiao (2001). The parameters, variables and indexes used are shown in Table 1. The objective is to minimise the total travel distance (and by that, the total time, observing speed limits and assuming different maximum speeds for salting and deadheading), for servicing all required arcs with the vehicle(s) in use, including deadheading:

$$\text{Minimize } \sum_{p=1}^k \sum_{i,j \in A} C_{ij} X_{ij}^p \quad (1)$$

The minimisation has to be carried out with respect to the following constraints. The vehicle capacity constraint ensures that the vehicle capacity is not exceeded:

$$\sum_{i,j \in A} L_{ij}^p Q_{ij} \leq W \quad \forall p = 1..k \quad (2)$$

The continuity constraint ensures that the vehicle enters and exits a node every time one is used, thus maintaining connectivity:

$$\sum_{\substack{r=1 \\ r \neq i}}^n X_{ri}^p - \sum_{\substack{r=1 \\ r \neq i}}^n X_{ir}^p = 0 \quad \forall i = 1..n \quad p = 1..k \quad (3)$$

The frequency constraint ensures the service of the demanded arcs. These arcs can either require service in one direction:

$$\sum_{p=1}^k L_{ij}^p + L_{ji}^p = 1 \quad \forall (i, j) \in A \quad (4)$$

...or a finite number of times, N_{ij} in either direction if the width of the carriageways exceeds the salting width¹:

$$\sum_{p=1}^k L_{ij}^p = \sum N_{ij} \quad \forall (i, j) \in A, N_{ij} \geq 1 \quad (5)$$

The service/deadheading constraint ensures that the service of an arc can only be carried out when a vehicle traverses the arc:

$$X_{ij}^p \geq L_{ij}^p \quad \forall (i, j) \in A, p = 1..k \quad (6)$$

The time constraint ensures that the total travel time is not exceeded:

$$\sum_{i,j \in A} T_{ij} X_{ij}^p + \sum_{i,j \in A} G_{ij} L_{ij}^p \leq T \quad \forall p = 1..k \quad (7)$$

The distance constraint ensures that the maximum total distance (service and dead-heading) is not exceeded:

$$\sum_{i,j \in A} X_{ij}^p C_{ij} \leq D \quad \forall p = 1..k \quad (8)$$

¹Based on the salting vehicles used in Denmark, we assume a salting width of 8 m.

The sub-tour cancellation constraint ensures the cancellation of sub-tours, which is a loop of tours without connection or inclusion of the depot, adapted from Golden and Wong (1981). When the network has at least one arc that requires service, then the right side of Eq. 9 will have a value greater than 0. To satisfy this equation, the left side containing the decision (flow) variables must be positive. According to Eq. 10, an arc's decision variable can only be positive if a vehicle travels the arc. Thereby, Eq. 9 prevents sub-tours by ensuring that at least one arc which is connected to the depot is used when traveling.

$$\sum_{\substack{r=1 \\ r \neq i}}^n F_{ir}^p - \sum_{\substack{r=1 \\ r \neq i}}^n F_{ri}^p = \sum_{j=1}^n L_{ij}^p \quad \forall i = 2..n, p = 1..k \quad (9)$$

$$F_{ij}^p \leq n^2 X_{ij}^p \quad \forall (i, j) \in A, p = 1..k \quad (10)$$

$$F_{ij}^p \geq 0 \quad \forall (i, j) \in A, p = 1..k \quad (11)$$

Due to the computational complexity of solving the equations shown above, a heuristic algorithm has to be developed that finds a near-optimal solution with respect to those constraints.

4 The Heuristic Algorithm

The heuristic proposed here is adapted from Resende and Ribeiro (1997) and based on the *Greedy Randomized Adaptive Search Procedure* (GRASP). GRASP consists of two phases: In the first phase, a feasible solution is generated by a *Route Construction Algorithm* (RCA) which is then improved in the second phase that searches for a better solution in its neighbourhood using a *Simulated Annealing* (SA) meta-heuristic to execute the local search for all the iterations. The aim of the iterative GRASP process is to reach the global minimum, i.e., to minimize overall salting time.

4.1 Phase One: Route Construction Algorithm

RCA traces a route from the depot by adding arcs incrementally with respect to the constraints (Sect. 3) for each iteration until a route is constructed. In every iteration of RCA, a list of arc candidates, a *Restricted Candidate List* (RCL), is generated by evaluating possible arc candidates that meet a greedy evaluation function. Furthermore, these arc candidates must be able to meet the feasibility of the partial solution, thus enabling arc exchange. This is used by the greedy evaluation function, which calculates the change in total cost and only considers the candidates that contribute with an incremental cost below the threshold value. The selection of candidates from

Table 2 The nomenclature of RCA

Variable	Definition
α	The threshold parameter between 0 and 1
C_{max}	Highest incremental cost
C_{min}	Lowest incremental cost
E	Possible candidate list
NO_{Routes}	The number of developed routes

RCL is random and repeated until a final route is obtained. The *cost* refers to the adjusted lengths, which are then adjusted for the traverse time based on the speed limit. In addition, the arcs with a service demand will be termed the required arcs.

The RCA consists of the following steps (see Table 2 for the nomenclature):

- Step 1:** The shortest distance between each pair of nodes is calculated and NO_{Routes} is set to 0. The shortest distance can be calculated using Floyd–Warshall’s algorithm (Cormen et al. 1990, pp. 558–565), which computes the shortest distances between all nodes in the graph.
- Step 2:** The current node is set to the depot node, partial route termed as no route, partial cost set to 0, the remaining capacity is set to the capacity of vehicle and the depot node is set as start node. If all the arcs’ requirements are 0, i.e. no arcs require service in the network, then the process goes to step 7, otherwise to step 3.
- Step 3:** The process goes to step 4 if any required arcs are connected to the current node, otherwise step 6.
- Step 4:** The RCL with a cost in the range of $R = [C_{min}, C_{min} + \alpha * (C_{max} - C_{min})]$ is generated. α is a value between 0 and 1.
- Step 5:** If the RCL contains at least one arc, then a random selection is made and the selected arcs are added to the partial route. Then, the partial route is updated by setting it to the current partial route including the added arc and partial cost is set as partial cost including the cost of the added arc. Furthermore, the remaining capacity is set to the remaining capacity minus the required capacity for servicing the added arc. If the added arc has a traverse demand ≥ 1 , then the demand is reduced by 1. In addition, the current node is set to end node of the recently added arc and the process returns to step 3, otherwise to step 6.
- Step 6:** If the addition of the required arc(s) is not possible due to violation of the constraints, then the shortest return route to the depot has to be determined. While doing this, the partial route is updated with the shortest return path to the depot, route cost is set to partial cost including the distance back to the depot and NO_{Routes} is set to $NO_{Routes} + 1$. After doing this, the process returns to step 2. Otherwise, the closest node connected to a required arc that does not exceed the constraints has to be determined and the shortest distance calculated.

Then, the partial route is updated to the partial route including the distance to the closest node, route cost is set to partial cost including the cost of the path to the closest node and the current node is set to the closest node and the process can go back to step 4.

Step 7: The solution is a set of routes that fulfills the demand of the required arcs in the network. NO_{Routes} is set to the total number of solution routes in the network.

4.2 Phase Two: Simulated Annealing Heuristic

The SA process starts with the best overall solution obtained from RCA and looks for a better one in a predefined neighbourhood. Since SA is a meta-heuristic, it allows for uphill moves, i.e. moves that generate solutions with higher costs than the current solution. This enables the process to escape local minima and find better solutions. The following pseudo-code outlines SA (Rere et al. 2015):

Algorithm 1 The local search

```

1: procedure SIMULATED ANNEALING
2:   Select the best solution vector  $x_0$  to be optimised. Initialise the parameters:
3:   Temperature  $T$ , Boltzmann's constant  $k$ , reduction factor  $c$ 
4:   while termination criterion is not satisfied do
5:     for number of new solution select a new solution:  $x_0 + \delta x$ 
6:       if  $f(x_0 + \delta x) < f(x_0)$  then  $f_{new} = f(x_0 + \delta x)$ ;  $x_0 = x_0 + \delta x$ 
7:       else  $\delta f = f(x_0 + \delta x) - f(x_0)$ 
8:         random  $r(0, 1)$ 
9:         if  $r < \exp(-\delta f/kT)$  then  $f_{new} = f(x_0 + \delta x)$ ,  $x_0 = x_0 + \delta x$ 
10:        else  $f_{new} = f(x_0)$ , do
11:          end if
12:        end if
13:         $f = f_{new}$ 
14:        Decrease the temperature periodically:  $T = c * T$ 
15:      end for
16:    end while
17: end procedure

```

The SA algorithm itself is embedded in a loop of moves of arcs in the route, followed by the execution of a route improvement algorithm, as outlined in the following.

The Movements

The movements for the local search are adapted from Pearn (1989), following the idea of “shuffling” subsets of arcs between two possible routes in order to find a better solution. This method randomly applies five arc selection rules to then select the optimal solution. This is done by defining sets of moves where each contains

Table 3 The nomenclature of the movements and the route improvement

Variable	Definition
$Level_1$	The obtained routes, when the input route is splitting at 2 points and rearranged
$Level_2$	The obtained routes, when the input route is splitting at 3 points and rearranged
N_1	The number of best $Level_1$ routes
N_2	The number of best $Level_2$ routes
$Routex_1$	The route from which one or more arcs are removed
$Routex_2$	The route in which one or more arcs are inserted

every move only once. The randomness is achieved by sequencing the order of the moves in the sets randomly. Each move of each set is performed for a number of finite iterations. When all iterations for all moves in one set have been performed, the process continuous to a different set. Table 3 shows the nomenclature used for the moves in the following steps.

Move 1 moves one arc from $Routex_1$ to $Routex_2$ with respect to the constraints of the succeeding route by executing the following four steps:

Step 1: First, a candidate list containing arcs that can be moved from one route to another, is generated. These arcs have to respect the constraints of the route to which they are moved to. For this move, there are following three categories of such arcs; **(a)** Both arcs being non-required arcs. **(b)** One of the preceding or succeeding arcs being a service arc. **(c)** Both arcs being service arcs.

This move can be executed with the preference hierarchy (probability of selection) of **(a)** followed by **(b)** and then **(c)**.

Step 2: For every candidate arc, all the candidate routes, in respect to the constraints, are determined.

Step 3: Then, an arc is randomly selected from the candidate list in respect to the probability of selection.

Step 4: For the selected arc, a candidate route is selected randomly, cf. step 2.

Move 2 exchanges a pair of arcs between $Routex_1$ and $Routex_2$ with respect to the constraints of both routes by executing the following two steps:

Step 1: First, a candidate list containing all pair of arcs that can be exchanged between two routes is generated by complete enumeration. This is done with respect to the constraints of both routes.

Step 2: Then, a random selection from the candidate list is made and the move is carried out by inserting the two arcs in their receptively counter route.

Move 3 exchanges two arcs from $Routex_1$ with one arc from $Routex_2$ with respect to the constraints of both routes by executing the same steps as in move 2.

Move 4 exchanges three arcs from $Route_{x_1}$ with one arc from $Route_{x_2}$ with respect to the constraints of both routes by executing the same steps as in move 2.

Move 5 exchanges three arcs from $Route_{x_1}$ with two arcs from $Route_{x_2}$ with respect to the constraints of both routes by executing the same steps as in move 2.

The Route Improvement

When a move is executed, a *Route Improvement* (RI) algorithm is used to generate a new route, starting and ending at the depot, including the moved arcs. The RI-algorithm uses the set of arcs obtained from the different moves and works with one set at a time. When the RI-algorithm has constructed a route, it splits the route at different points, creating multiple elements. Then, the algorithm changes the sequence of elements, by moving the elements, creating multiple routes with different arranged sequence of elements. Among these routes, the cheapest route with rearranged sequence of elements is the final result of the RI-algorithm, which then is used in the SA as the result of the move. The steps of the RI-algorithm are as follows;

- Step 1:** The RCA, as described in Sect. 4.1, is used to create a route that starts and ends at the depot.
- Step 2:** The route obtained from RCA is then divided in all possible ways at two points creating three segments ($Level_1$ -analysis). The sequence of these three segments is then changed and rearranged in all possible ways with respect to the orientation of the arcs. This step ends with an evaluation of the cost of each route after the rearrangement.
- Step 3:** The N_1 routes of the $Level_1$ -analysis are stored in a $Level_1$ -list. Every time a cheaper route than N_1 is found, the list is updated by including the new route.
- Step 4:** The top N_1 routes are then randomly selected from the $Level_1$ -list, termed $Level_2$ -routes, and stored as a $Level_2$ -list. Each route from this list is then divided in all possible ways at three points creating four segments, $Level_2$ -analysis. The sequence of these four segments are then changed and rearranged as in step 2.
- Step 5:** If the $Level_2$ -analysis creates a cheaper route, N_2 , than the routes in the $Level_1$ -list, then the N_2 -route is used in the $Level_1$ -analysis, step 2. If the new $Level_1$ -analysis of the N_2 -route creates a cheaper route, then a new $Level_2$ -analysis is generated, step 4.
- Step 6:** The final result is the cheapest route after the performance of $Level_1$ - and $Level_2$ -analysis.

The low cost routes for both $Route_{x_1}$ and $Route_{x_2}$ are thereby determined. After creating the sets of moves, the SA can be performed, consisting of the following steps and using the nomenclature from Table 4:

- Step 1:** The values of the parameters; a , N_{max} , N_{move} , $Max_{Iteration}$ and T have to be set. In addition, the following parameters have to be; $N = 0$ and $N_{iteration} = 0$. The process continuous to step 2.
The cooling of SA is managed by N_{max} and a i.e. for every N_{max} , the

Table 4 The nomenclature of SA

Variable	Definition
a	The temperature reduction factor
$Best_{Cost}$	The lowest cost
$Best_{Routes}$	The corresponding route to the lowest cost
$Curr_{Cost}$	The current cost
$Curr_{Routes}$	The set of corresponding routes to the current cost
$Max_{Iteration}$	The maximum number of iterations
N	The number of iterations at the current temperature (counter)
$N_{Iteration}$	The number of ongoing iterations (counter)
N_{max}	The maximum number of iterations at each temperature
N_{move}	The maximum number of iterations for each type of move
R	A random value between 0 and 1
RCA_{Cost}	The total cost obtained from RCA
RCA_{Routes}	The set of routes obtained from RCA
SA_{Cost}	The cost of SA
SA_{Routes}	The set of SA routes
T	The temperature
ΔTC	The value of $(Curr_{Cost} - SA_{Cost})$

temperature reduces by factor a . In addition, the temperature T allows the exchange of the neighbouring solutions, enabling the heuristic search space, thus the value of T has to be high.

Furthermore, in this step, a set of moves has to be generated, as described in Sect. 4.2.

Step 2: The following conditions have to be set; $Best_{Cost} = RCA_{Cost}$, $Best_{Routes} = RCA_{Routes}$, $Curr_{Cost} = RCA_{Cost}$ and $Curr_{Routes} = RCA_{Routes}$. The process continuous to step 3.

Here, the route(s) and solution values obtained from the RCA are assumed and set to be the current and the best route with the related values.

Step 3: The following condition is set; $N_{iteration} = N_{iteration} + 1$. The process continuous to step 4.

In this step, the overall iteration counter increases by 1.

Step 4: The following condition is set; $N = N + 1$. The process continuous to step 5.

In this step, the annealing iteration counter increases by 1.

Step 5: The following conditions are set; SA_{Cost} = the cost of the performed move and SA_{Routes} = the route obtained after performing the move. The process continuous to step 6.

In this step, the cost and set of route(s) of the move are set as the SA values.

- Step 6:** $\Delta TC = Curr_{Cost} - SA_{Cost}$ is calculated and depending on the result, the process continuous to either step 7 or step 8.
In order to assess the solution obtained after a move, the difference between the current cost and the SA cost has to be calculated.
- Step 7:** The following conditions are set if $\Delta TC \geq 0$: $Curr_{Cost} = SA_{Cost}$ & $Curr_{Routes} = SA_{Routes}$ and the following conditions are set if $SA_{Cost} < Best_{Cost}$, $Best_{Cost} = SA_{Cost}$ and $Best_{Routes} = SA_{Routes}$. The process continuous to step 9.
Here, based on the evaluation of the difference between the current cost and the SA cost a number of conditions are set. In case of the difference being greater or equal then 0, the solution of SA is better the current solution. At the same time, the values of SA are set as the best values, if the cost of SA is less than the cost of the best solution.
- Step 8:** The following conditions are set in case of $\Delta TC < 0$ and $\exp(\Delta TC/T) > R$; $Curr_{Cost} = SA_{Cost}$ and $Curr_{Routes} = SA_{Routes}$. The process continuous to step 9.
In contrast to step 7, this step deals with current solutions being worse than the SA solutions. If so, the non-improving SA solutions have to, based on the probability function, either be accepted or rejected. As $\Delta TC \mapsto 0$, the chance of the non-improving SA solutions being accepted increases. At the same time, as $T \mapsto 0$ the probability of acceptance decreases.
- Step 9:** The following conditions are set if $N = N_{max}$; $T = a * T$ and $N = 0$. The process continuous to step 10.
If the maximum number of iterations at the current temperature has been reached, the temperature is reduced by a and the N counter is set to 0
- Step 10:** If $N_{Iteration} = Max_{Iteration}$, the algorithm process ends. Otherwise, the process continuous to step 11.
If the total number of ongoing iterations has reached the maximum number of iterations (the algorithm's stop condition), then the best overall solution of the iterations is the final solution.
- Step 11:** If $N_{Iteration} \text{ MOD } (5 * N_{move}) = 0$, then a new set of moves has to be generated and the process returns to step 3. Otherwise, the process continuous to step 12.
If the number of ongoing iterations has reached five times the maximum iterations for each type of move, then all solutions for that set of moves have been searched. Since the $N_{iteration}$ has not reached the number of maximum iteration of the algorithm, then a new set of moves can be assessed.
- Step 12:** If $N_{Iteration} \text{ MOD } N_{move} = 0$, then the type of move, in the same set of moves, has to be updated and the process returns to step 3. If that is not the case, the process still returns to step 3.
Here, it is determined whether there has to be a change in the type of move or if the algorithm has to run the next iteration of the same type of move.

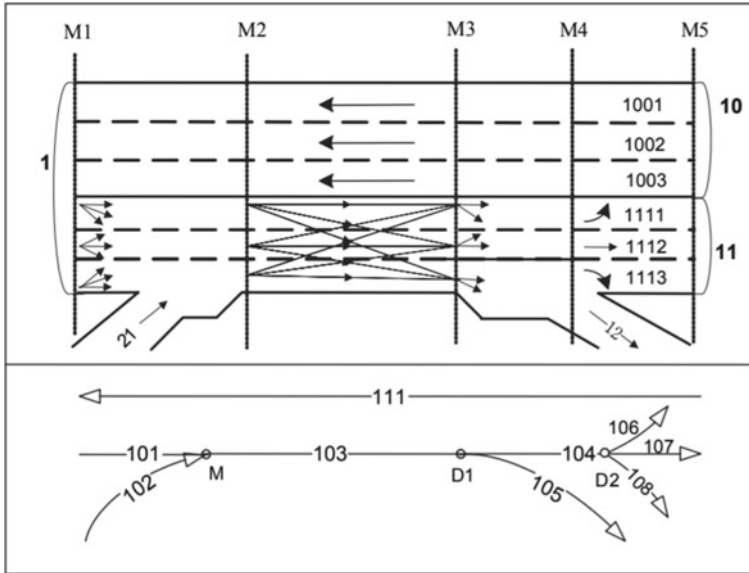


Fig. 2 An example of a traffic direction-based model (Ruas and Gold 2008, p. 623)

The best overall solution becomes the final solution when one of the stop conditions has been reached.

5 Implementation and Results

In order to implement and test the approach outlined in the previous sections, three of VD's winter maintenance routes have been modeled based on a traffic direction-based schema. This model is an abstraction of vehicle driving directions rather than physical objects (Ruas and Gold 2008, p. 622). Figure 2, shows an example in which the northern carriageway is represented as a single line due to same direction lanes and the southbound carriageway is represented by five modeling objects based on their traffic directions; off-ramp, on-ramp, right-turn, left-turn and ahead flow.

The data for the three routes (and their vicinity; see Fig. 3 for an example) were obtained from VD's road management system² and, after extensive cleaning and manual fixing of disconnected arcs, translated to the traffic direction-based schema. RCA and SA have then been implemented in Python to work on this model.

The three test routes used for evaluation of the approach are shown in Figs. 1 and 4, respectively. After some experiments, the optimization procedure was run with an initial temperature T of 100, a temperature reduction factor α of 0.995, a maximum

²See <http://vejman.dk> and <https://trafikkort.vejdirektoratet.dk>.

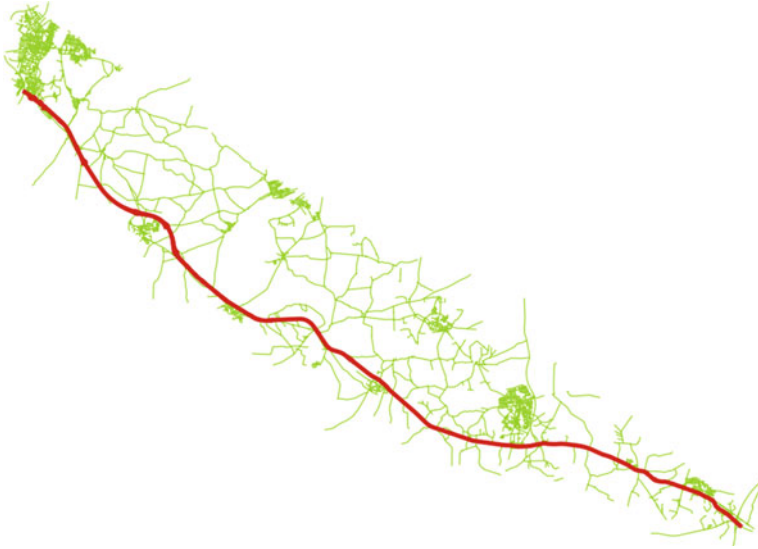


Fig. 3 A maintenance route with neighbourhood, obtained by clipping a buffered convex hull



Fig. 4 Winter maintenance routes K18 and K20

number of iterations for each type of move N_{move} of 1000, a maximum number of iterations at each temperature of 300, and an overall maximum number of iterations of 125.000. The SA progress with these parameters for each route is shown in Fig. 5. In addition to a fixed salting width of 8 m and the arcs' respective speed limits, the evaluation is based on a maximum speed of 80 km/h when deadheading, 30 km/h when salt spreading the ramps and 70 km/h when salting elsewhere.

The optimization produces significantly shorter routes for all three routes, despite the fact that the maintenance route itself was fixed and optimization was only possible in the deadheading and in the order of the segments of the maintenance route. For test route K18, the route length was reduced by 6.3% and service time by 7.1%. For K20, route length was reduced by 8.2% and service time by 10.7%, and for K72, route length was reduced by 11.7% and service time by 10.6%. Assuming the current

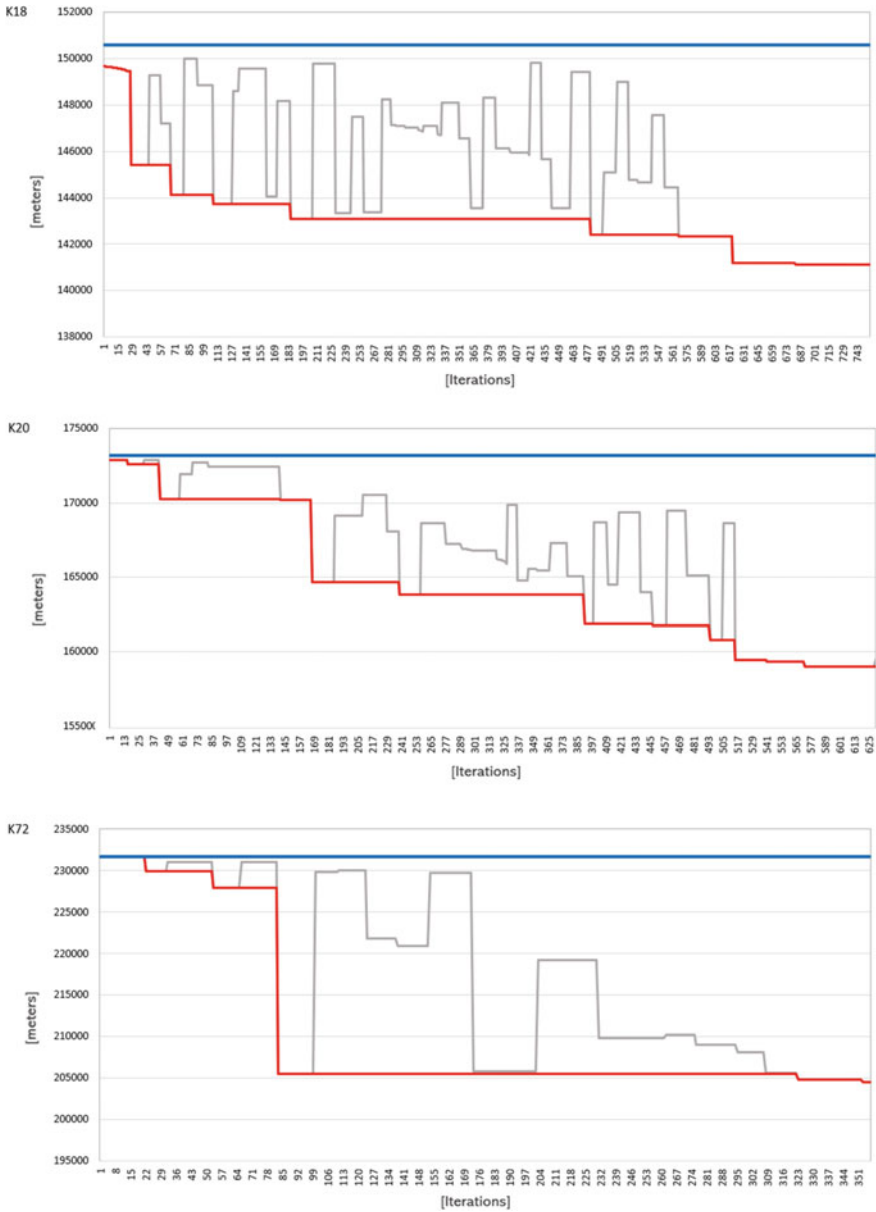


Fig. 5 SA progress for the three test routes, with the blue line indicating the route length before optimization, the gray line indicating the route length at the current iteration, and the red line indicating the minimum route length found up to the current iteration

average salting cost per hour and that similar improvements can be expected for the remaining roads VD has to maintain during the winter, we expect savings in the order of 2.8 million DKK (about 380.000 Euros). This estimate does not include savings due to the automation of route planning.

Concerning the route optimization procedure itself, there is still room for improvement in computation time. The result shown here were obtained on a standard laptop and took 7 h 28 min (K18), 12 h 41 min (K20), and 71 h 3 min (K72). While these are clearly substantial computation times, they are not unusual for such heuristic procedures on complex problems. The computation time of almost three full days for route K72 is a prime example, as this is the only route that requires two vehicles to be serviced, resulting in a substantially increased processing time. However, the focus of this research has been on demonstrating that substantial improvements of the routes themselves are possible using this approach, without a big focus on computation times, since these procedures are only run once in the rare event that the routes are changed. Nonetheless, we believe that further testing and tuning of the parameters would allow us to improve computation times.

6 Conclusions

This paper has shown how the Danish Road Directorate's salt spreading can be optimized by treating it as a Capacitated Arc Routing problem (CARP). Based on CARPs categorisation as NP-hard, heuristic procedures were investigated and a heuristic algorithm based on a Greedy Randomized Adaptive Search Procedure (GRASP) was developed. In the proposed procedure, each iteration constructs a solution which is further optimised by performing a local search with a Simulated Annealing (SA) based heuristic algorithm. This is done by constructing an initial solution with a Route Construction Algorithm (RCA), that incrementally traces a route with the depot as the start and end node, in order to service the required arcs with respect to a number of constraints and criteria at each iteration. The final obtained solution of RCA is used to perform a Simulated Annealing (SA) based local search, which improves the RCA-obtained solution by exchanging arcs in terms of five different moves that define the solution space. When a move is performed, a Route Improvement (RI) algorithm is used to rearrange and reconstruct a new solution that, if better than the initial solution, is used as the new initial solution for the next iteration. The final result becomes the best overall solution that is obtained when the GRASP has reached one of the stop conditions. The proposed algorithms were implemented and tested on three different service routes, showing that they were able to improve the existing routing of all three test routes significantly. Future work on this approach should focus on the reduction of the processing time, as well as the inclusion of further parameters in the optimization, such as variable salting widths or the inclusion of typical morning traffic conditions, when most salting activities take place.

References

- Arakaki RK, Usberti FL (2018) Hybrid genetic algorithm for the open capacitated arc routing problem. *Comput Oper Res* 90:221–231. <https://doi.org/10.1016/j.cor.2017.09.020>, <http://www.sciencedirect.com/science/article/pii/S0305054817302502>
- Bodin LD, Golden BL, Assad AA, Ball MO (1981) The state of the art in the routing and scheduling of vehicles and crews. US Department of Transportation, Urban Mass Transportation Administration. <https://babel.hathitrust.org/cgi/pt?id=ien.35556021333117;view=1up;seq=13>
- Chapleau L, Ferland JA, Lapalme G, Rousseau JM (1984) A parallel insert method for the capacitated arc routing problem. *Oper Res Lett* 3(2):95–99. Accessed 04 Dec 2017
- Christofides N (1973) The optimum traversal of a graph. *Omega* 1(6):719–732. Accessed 03 Dec 2017
- Cormen TH, Leiserson CE, Rivest RL (1990) Introduction to algorithms, 1st edn. MIT Press and McGraw-Hill
- Eglese R (1994) Routeing winter gritting vehicles. *Discret Appl Math* 48(3):231–244. Accessed: 08 Dec 2017
- Eiselt HA, Gendreau M, Laporte G (1995) Arc routing problems, part ii: the rural postman problem. *Oper Res* 43(3):399–414. Accessed 15 Dec 2017
- Evans J, Weant M (1990) Strategic planning for snow and ice control vehicles using computer-based routing software. *Public Works* 121(4):60–64. Accessed 09 Dec 2017
- Gendreau M, Hertz A, Laporte G (1992) New insertion and postoptimization procedures for the traveling salesman problem. *Oper Res* 40(6):1086–1094. Accessed 13 Nov 2017
- Golden BL, Wong RT (1981) Capacitated arc routing problems. *Networks* 11(3):305–315. Accessed 15 Dec 2017
- Goode L, Nantung T (1995) CASPER: the friendly, efficient snow routes planner. <http://onlinepubs.trb.org/onlinepubs/trnews/rpo/rpo.tm181.pdf>
- Haghani A, Qiao H (2001) Decision support system for snow emergency vehicle routing: algorithms and application. *Transp Res Rec: J Transp Res Board* 1771:172–178. Accessed 02 Feb 2017
- Haslam E, Wright JR (1991) Application of routing technologies to rural snow and ice control. *Transp Res Board* 1304:202–211. Accessed 10 Sept 2017
- Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. *Science* 220(4598):671–680. Accessed 29 Nov 2017
- Knudsen F, Eram MM, Jansen KB (2014) Denmark, Technical Committee 2.4, pp 52–60
- Kurniawan R, Sulistiyono MD, Wulandari GS (2015) Genetic algorithm for capacitated vehicle routing problem with considering traffic density. In: 2015 international conference on information technology systems and innovation (ICITSI), pp 1–6
- Lemieux PF, Gampagna L (1984) The snow ploughing problem solved by a graph theory algorithm. *Civ Eng Syst* 1(6):337–341. Accessed 29 Nov 2017
- Male JW, Liebman JC, Orloff CS (1977) An improvement of Orloff's general routing problem. *Networks* 7(1):89–92. Accessed 11 Dec 2017
- Metropolis N, Rodenbluth AW, Rosenbluth MN, Teller AH (1953) Equation of state calculations by fast computing machines. *J Chem Phys* 21(6):1087–1092. Accessed 02 Feb 2018
- Nemhauser GL, Wolsey LA (1988) Integer programming and combinatorial optimization. Wiley, Chichester; Nemhauser GL, Savelsbergh MWP, Sigismondi GS (1992) Constraint classification for mixed integer programming formulations. *COAL Bull* 20:8–12
- Pearn WL (1989) Approximate solutions for the capacitated arc routing problem. *Comput Oper Res* 16(6):589–600. Accessed 03 Dec 2017
- Rere LR, Fanany MI, Arymurthy AM (2015) Simulated annealing algorithm for deep learning. *Procedia Comput Sci* 72(1):137–144. Accessed 23 Nov 2017
- Resende MGC, Ribeiro CC (1997) A grasp for graph planarization. *Networks* 29(3):173–189. Accessed 16 Oct 2017
- Ruas A, Gold C (2008) Headway in spatial data handling: 13th international symposium on spatial data handling. Springer

Part III
Geoinformation Science and Geospatial
Technologies in Urban/Regional Planning

Tracing Tourism Geographies with Google Trends: A Dutch Case Study



Andrea Ballatore, Simon Scheider and Bas Spierings

Abstract Search engines make information about places available to billions of users, who explore geographic information for a variety of purposes. The aggregated, large-scale search behavioural statistics provided by Google Trends can provide new knowledge about the spatial and temporal variation in interest in places. Such search data can provide useful knowledge for tourism management, especially in relation to the current crisis of tourist (over)crowding, capturing intense spatial concentrations of interest. Taking the Amsterdam metropolitan area as a case study and Google Trends as a data source, this article studies the spatial and temporal variation in interest in places at multiple scales, from 2007 to 2017. First, we analyze the global interest in the Netherlands and Amsterdam, comparing it with hotel visit data. Second, we compare interest in municipalities, and observe changes within the same municipalities. This interdisciplinary study shows how search data can trace new geographies between the interest origin (what place users search from) and the interest destination (what place users search for), with potential applications to tourism management and cognate disciplines.

Keywords Interest geography · Place search · Web science · Google Trends · Tourism · Amsterdam · Netherlands

A. Ballatore (✉)

Department of Geography, Birkbeck, University of London, London, UK
e-mail: a.ballatore@bbk.ac.uk

S. Scheider · B. Spierings

Human Geography and Planning, Utrecht University, Utrecht, The Netherlands
e-mail: s.scheider@uu.nl

B. Spieringsat

e-mail: b.spierings@uu.nl

© Springer Nature Switzerland AG 2020

P. Kyriakidis et al. (eds.), *Geospatial Technologies for Local and Regional Development*, Lecture Notes in Geoinformation and Cartography,
https://doi.org/10.1007/978-3-030-14745-7_9

1 Introduction

In their effort to provide relevant information to users, search engines create connections between users and Web content. Just considering Google, more than a billion people query the search engine every day, resulting in more than 3.5 billion requests per day.¹ Geography plays a central role in this process of information retrieval and consumption, as users, machines, and online resources are always located somewhere on the planet (Graham et al. 2015). Users search for places from their location for a variety of purposes, including travel, reference, news, entertainment, and investment. The expression of interest in an entity (e.g. typing “London” into Google) is a relatively novel and unique type of big data that can be used for the study of human dynamics (Stephens-Davidowitz 2013).

Tourism constitutes a major and increasing driver for the development of places. Over the past decade, global tourist flows have strongly increased, exhibiting a growth rate of 4% or more every year. In 2017, the growth was 7%, reaching a total of 1.32 billion overnight visitors worldwide.² Search engines, combined with online platforms such as booking.com and Airbnb, have radically transformed how tourists find their destinations and plan their journeys. The economic and cultural benefits of growing tourism do not come without a cost. Many focal areas in the world suffer from overcrowding, including Venice, Barcelona, and Amsterdam. In this context, big data analytics offer new insights into observing, modelling, and forecasting tourism-related behaviour in space and time (Önder and Gunter 2016).

As part of the broad “Web science” approach (Hendler et al. 2008), the potential of search data has attracted attention from multiple viewpoints. Since 2006, Google Trends³ has provided aggregated measures of interest in topics searched for on the search engine, capturing how interest in a person, city, event, book, team, or movie can rise, fall, concentrate, and disperse over time (Jun et al. 2018). This data source spurred several applications in economics (Choi and Varian 2012) and epidemiology (Lazer et al. 2014). While some of these works include geographic and temporal dimensions, no study has focused on the search for places with an explicit attention to the spatial dimension of this human behaviour.

In this article, we use Google Trends data to observe the variation of interest in places, with particular attention to potential applications in tourism studies. This exploratory study of what we might call an “interest geography” aims at paving the way towards more modelling-oriented efforts. As a case study, we consider the Amsterdam metropolitan area, a region of the Netherlands that comprises 33 municipalities. The investigation is restricted to the most recent decade with complete data (2007–2017). Central to the study are the interest origin (which place people search from) and the interest destination (which place people search for). Following

¹<https://web.archive.org/web/20181204153621/https://www.internetlivestats.com/google-search-statistics>. All URLs were accessed in November 2018, and are stored in the Internet Archive.

²https://web.archive.org/web/20180808124718/http://cf.cdn.unwto.org/sites/all/files/pdf/unwto_barom18_01_january_excerpt_hr.pdf

³<https://web.archive.org/web/20181206023025/https://trends.google.com/trends/>

recommendations by Singleton et al. (2016), all analyses in this article are fully reproducible, and code and data are available online.⁴ This article is a first step towards answering the following research questions:

- RQ1** How does search behaviour change at different geographic scales (e.g. national, regional, local)?
- RQ2** How does search behaviour change in space across different interest origins and different interest destinations?
- RQ3** Is there any correlation between search behaviour and measurable tourism activity, such as hotel visits?
- RQ4** What areas are over- and under-represented in online search interest with respect to population size?

The remainder of this article is organized as follows. Section 2 discusses related work in spatio-temporal analytics of search data, particularly on Google Trends, and tourism studies. The analysis of interest origin for the Netherlands and Amsterdam at the global scale is then discussed in Sect. 3, contrasting hotel visits with search interest. Section 4 analyzes the variation of interest in municipalities in the Amsterdam metropolitan area over time, both within each municipality and between municipalities. Section 5 concludes the article with a discussion and outlook on the potential of search data for tourism analytics.

2 Related Work: Search Interest and Tourism

This study is located at the intersection of web science, internet geography, data science, and tourism studies. This section aims at covering relevant works from these areas, including background about tourism in the Amsterdam metropolitan area.

2.1 Search Engine Data for Human Dynamics

The information retrieval community has traditionally relied on search engine logs as a source of information about user behaviour, before privacy concerns emerged during the notorious AOL incident in 2006.⁵ Since then, search data has been released only in coarse aggregated forms, making it hard to connect searches with specific individuals. Additionally, as the business model of Google relies on the secrecy of its algorithm to reduce ranking manipulation in the arena of Search Engine Optimization

⁴<https://github.com/andrea-ballatore/SearchGeography>.

⁵<https://web.archive.org/web/20181204192954/https://www.nytimes.com/2006/08/09/technology/09aol.html>.

(SEO), access to highly granular search data is essentially impossible outside of the search engine provider.

For these reasons, Google Trends indicates an index of interest from 0 to 100, calculated relative to *all searches* conducted in a given period, allowing comparisons either between terms, or within the same term over time, and does not provide the actual number of searches. Therefore, if the overall pool of searches grows and changes composition, the index for a given topic can *decrease* even if the actual number of searches has increased. Given the noisy, volatile, and ambiguous nature of search data, caution is needed when handling the Google Trends index.⁶

Search motivations vary widely (e.g. searching for London to plan a journey or to learn about its history), and the disaggregation of these behaviours is not trivial. Search terms are often ambiguous (“Chelsea” as an English or American neighbourhood, an English football team, or a female name), and exceptional events, such as natural and man-made catastrophes, tend to generate short-lived bursts of interest. Spatially, the geography of searches is biased by an uneven user distribution: Google has a limited user base in major countries such as China, Russia, and Iran. Furthermore, virtual private networks (VPNs) deliberately obfuscate the geographic location of users.

Despite these limitations and challenges, Google Trends data have been used in a variety of contexts, primarily in economics (Jun et al. 2018) and management (Askitas and Zimmermann 2015). Researchers showed that search data can forecast future behaviours (Goel et al. 2010), often correlate with changes in the value of several goods and services (Choi and Varian 2012), and can quantify attitudes that are not easily revealed by polls, such as racism (Stephens-Davidowitz 2013). From a more geographical perspective, Google Flu Trends famously linked search locations to potential flu outbreaks, producing some hype and subsequent reappraisal of the potential of big search data for forecasting (Lazer et al. 2014).

In GIScience and geography, our prior work focuses on several facets of search engines, including the localness of results (Ballatore et al. 2017), interest in crowd-sourced datasets (Ballatore and Jokar Arsanjani 2018), and popularity of data science tools (Ballatore et al. 2018). However, none of these studies has the spatial focus that we take in this article.

2.2 *Tourist (Over)Crowding and Big Data*

Cities and their historical cores have since long been popular recreational destinations among travellers (Gospodini 2006), but only since the end of the 20th century assets are purposefully developed and marketed in order to create and foster a “visitor economy” (Hall and Barrett 2017). Ambitious city marketing strategies were often developed and opportunistic city development projects implemented to sell

⁶<https://web.archive.org/web/20181108133214/https://medium.com/@pewresearch/using-google-trends-data-for-research-here-are-6-questions-to-ask-a7097f5fb526>.

and boost the city's appeal for visitors in terms of leisure and consumption, with the goal of attracting large numbers of visitors and their spending power (Spierings 2013). Supported by, amongst others, the rise in leisure time, the growth of spending power, and the proliferation of low-cost airlines, many cities have also become very successful in attracting tourists. However, rising complaints and protests by residents against large and increasing numbers of tourists are clear indications that several cities have recently become *too* successful, and are increasingly perceived by residents and tourists alike as “(over)crowded” (Popp 2012; Novy and Colomb 2017; Neuts and Vanneste 2018).

Even though surveys provide valuable quantitative data about tourist flows and related processes of (over)crowding, supporting decision-making, the rise of “big data” promises cheaper and more granular data sources (Kitchin 2014). The potential of search engine behaviour has not gone unnoticed in tourism and management studies (Pan et al. 2012; Yang et al. 2015). Search statistics can provide valuable information about the interests and intentions of tourists (Li et al. 2017). The estimation of tourist arrivals can also be enriched by search trends, as shown in a case study on Vienna (Önder and Gunter 2016). Search interest correlates better with arrivals when linguistic and other biases are taken into account (Dergiades et al. 2018). Along similar lines, Siliverstovs and Wochner (2018) improved the forecasting accuracy of Google trends for tourist arrivals to Switzerland by employing a cross-sectional instead of a longitudinal approach. This body of work offers sophisticated insights into the temporal dimension of search patterns, but overlooks the spatial dimension that we focus on in this article.

2.3 Tourist Marketing and Spreading in Amsterdam

As our case study focuses on the Amsterdam metropolitan area, it is beneficial to provide some background about the city's complex relationship with tourism. Nowadays, most tourist destinations aim to manage and mitigate the social impacts of (over)crowding. One of the strategies that is experimented with is the spatial “spreading” of tourists. Amsterdam is a telling case of this policy strategy, which has the objective of spreading tourists at the urban, regional, and even at the national level. Recently, a marketing campaign was initiated to promote neighbourhoods outside and adjacent to the historical core of the city as interesting and diverse destinations.⁷

A complementary tourist-spreading strategy was developed at the national level with Amsterdam as an important constituent of several marketing “storylines”. These narratives present the country as one metropolis (i.e. Holland city) and promote visits “off the beaten track” throughout the country.⁸ The storyline of the Dutch Golden

⁷<https://web.archive.org/web/20181206203125/https://www.iamsterdam.com/nl/over-ons/amsterdam-marketing/afdelingen/marketing-strategy/consumer/buurtencampagne>.

⁸<https://web.archive.org/web/20181206203356/https://www.nbtc.nl/en/homepage/collaboration/storylines.htm>.

Age, for instance, combines the well-known capital's canals and the Rijksmuseum with attractions in cities like Haarlem, Leiden, Middelburg, and Dordrecht, inviting tourists to go beyond Amsterdam.

Another marketing campaign at the regional level has been running for more than a decade.⁹ Several places and sites within the urban region are marketed as extensions of Amsterdam—including Zandvoort aan Zee (a seaside resort) as part of “Amsterdam beach”, Muider slot (a medieval castle in Muiden) as “Amsterdam castle”, Zaanse Schans and Volendam (respectively a traditional village with windmills and a traditional fishing village) as “old Holland”, and Almere (a city developed in the second half of the twentieth century) as “new land”. Our study of search geography brings a new viewpoint on how to sense and gather data about the spatial configuration of tourist interest.

3 Interest Origin for the Netherlands and Amsterdam

To inspect the geography of search, we start by observing the interest origin (i.e. where people search from) for the Netherlands and Amsterdam, two prominent interest targets in our Dutch case study. Hence, we collected the Google Trends index (GTI) from 2007 to 2017 at the country level.¹⁰ For example, the GTI for Amsterdam in France in 2010 is 2, with 100 being the GTI in the country that showed the highest interest (the Netherlands). In order to gain an understanding of how many searches actually occurred, we obtained estimates from SEO company SemRush. The company indicates that, in October 2018, “Amsterdam” received 246,000 searches from the Netherlands and 201,000 from the UK. These estimates are useful to sense the magnitude of the search volume.

During a first inspection of the GTI, we noted an unusual peak in July 2014 for “Netherlands”, corresponding to the World Cup. It is therefore highly likely that most of those searches were aimed at the Dutch national football team and not at the country. “Amsterdam”, by contrast, showed seasonal variation without bursts. Hence, in the most of the analyses below we focused on Amsterdam as opposed to the Netherlands, for which correction would be needed. To reduce the noise in the data, unless specified otherwise, we retrieved Google Trends data with the “low volume” option off. This option allows to include or exclude data points with low values that are filtered out by default.

In Google Trends, it is possible to retrieve data as a “term” (a plain string), or as a “concept” (a bundle of correlated terms). While, in principle, the results can vary significantly between terms and concepts, empirical observation indicated that the variation was negligible in this study, and therefore we always used terms.

⁹<https://web.archive.org/web/20181206102019/https://www.iamsterdam.com/en/plan-your-trip/day-trips>.

¹⁰The data was retrieved in tabular form from the Google Trends API using the R package *gtrendsR*: <https://cran.r-project.org/web/packages/gtrendsR>.

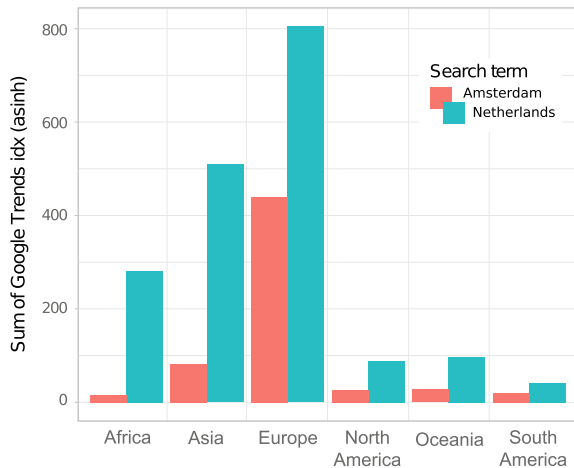
As the difference in interest for objects of different sizes can vary by orders of magnitude, when appropriate, we show transformed values with inverse hyperbolic sine, conceptually similar to a logarithmic transformation, but better at handling 0 values.

3.1 Spatial Distribution of Interest Origin

The GTI quantifies the interest for the interest destinations (Amsterdam and the Netherlands) in each country in the time period, relative to other countries, which makes the comparison between countries possible. As a first step to describe the overall geography of interest origin, we aggregated the countries by continent (omitting Antarctica), summing the GTI for the 11 years included in the dataset, as shown in Fig. 1. For all continents, the Netherlands obtained higher interest than Amsterdam, suggesting a proportionality between the size of the interest target and the GTI (RQ1). Europe shows the highest GTI for both terms by far, with the Netherlands obtaining 1.7 times the summed GTI than Amsterdam. Notably, Africa obtained 17.6 times more interest in the country than in the capital, while the same ratio for Asia is 6.2. In this sense, searches from Africa focus disproportionately on the whole country as opposed to the capital city. The other continents show substantially lower values, indicating a strong distance decay effect. At a closer look at European countries, the Netherlands generates the highest GTI (100) for both terms, followed by Belgium, UK, Ireland, Germany, and Italy.

At a finer spatial granularity, Fig. 2 shows the variation in GTI by country for Amsterdam between 2007 and 2017 (RQ1). Note that for this comparison, low-volume data was included, as otherwise most African countries would have been missing. These maps clearly show that interest origin became less diverse and more

Fig. 1 Google Trends interest (GTI) for terms “Amsterdam” and “Netherlands”, summing countries by continent from 2007 to 2017 (GTI transformed with inverse hyperbolic sine, without low-volume data)



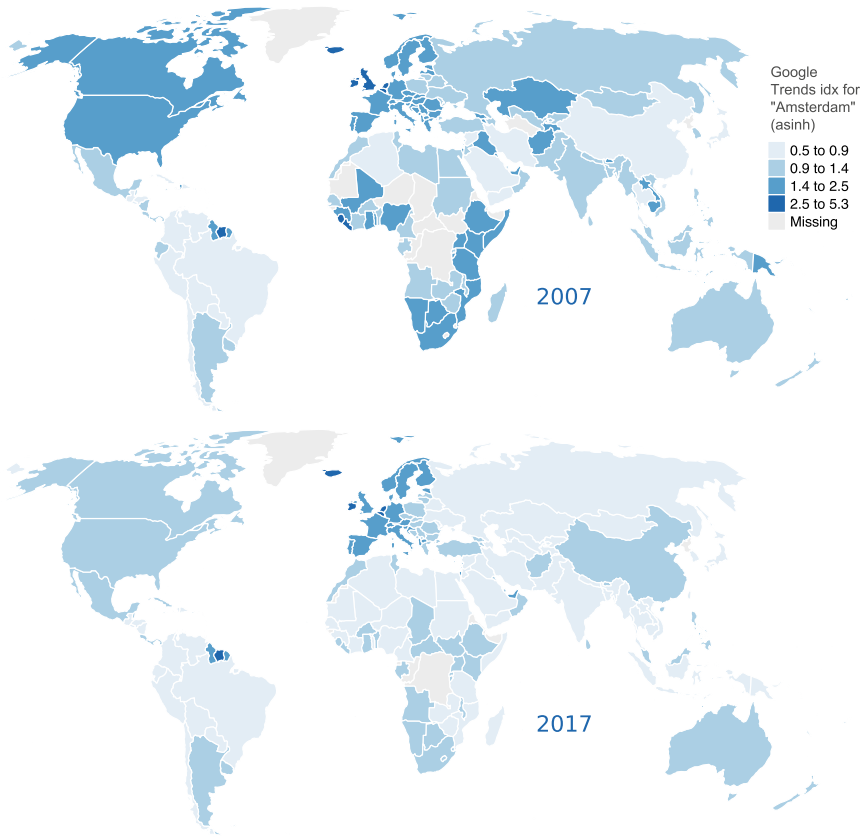


Fig. 2 Google Trends interest for terms “Amsterdam” by country 2007 and 2017, with low-volume data (index transformed with inverse hyperbolic sine). Data grouped with quantiles. Period 2007–2017. Projection: Robinson. World borders from Natural Earth, 2017

concentrated in Europe over time (RQ2). A possible explanation for this change is that, in the decade being analyzed, European tourist flows grew comparatively more than those from other continents (see Sect. 3.2). Beyond the striking dominance of Europe, post-colonial links emerged. Countries with unusually high interest (GTI in range [8,23]) include former or current Dutch territories, such as Suriname, Aruba, Curaçao, and Sint Maarten. By contrast, Liberia also has an unusually high GTI, whereas colonial relations with the Netherlands were very short-lived.

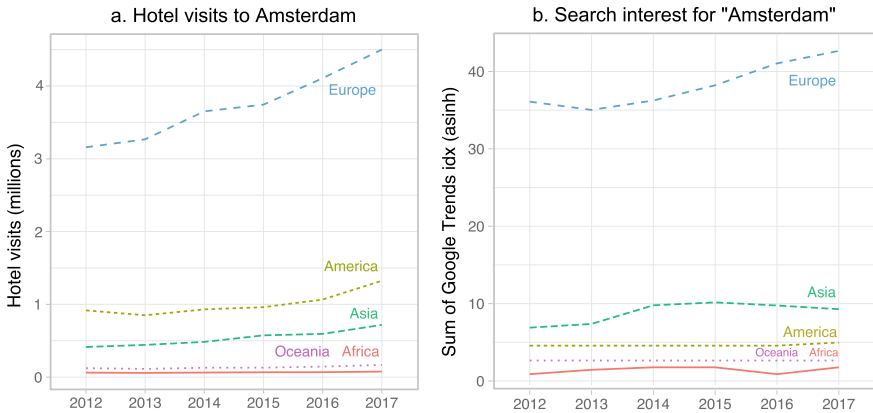


Fig. 3 **a** Hotel visits in Amsterdam and **b** GTI for term “Amsterdam” from 2012 to 2017, without low-volume data (index transformed with inverse hyperbolic sine) for 50 countries. Sources: Statistics Netherlands and Google Trends

3.2 Temporal Change in Interest and Hotel Visits

To study the relationship between quantifiable tourist behaviour and search data (RQ3), we retrieved hotel visits in Amsterdam from 2012 to 2017 for 50 countries.¹¹ This data estimates stays in traditional hotels, and does not include home sharing platforms, such as Airbnb. Figure 3 depicts the yearly change in hotel visits and the search interest as the sum of transformed GTI per continent in the same period, for the 50 countries included in the dataset. It is possible to note that a superficial similarity between the trend lines for the case of Europe, dominant and increasing both in terms of hotel visits and GTI. However, the trends diverge for other continents, which show a slight decline in GTI, while hotel visits actually increased, even though not at the same pace as for Europe (RQ3).

The number of years in the dataset (6) is too low for meaningful correlation analysis at the country level, we calculated the pair-wise correlation of all countries for all years. The hotel visits are highly granular (181 levels in the variable), but because of the coarse granularity of GTI (the variable takes only 12 levels in the case of Amsterdam), the pairs contain a large number of ties, and therefore the non-parametric Kendall’s τ is a suitable correlation coefficient. Table 1 shows the global pair-wise correlation between hotel visits and GTI. No meaningful correlation is found for the Netherlands (τ near 0), confirming the difficulty of extracting tourism-related value from Google Trends with a query affected by interest bursts (e.g. the World Cup). By contrast, a strong positive correlation is found for Amsterdam, without low-volume data ($\tau = .57$).

Over time, this correlation appears to have decreased (from .63 in 2012 to .51 in 2017, at $p < 0.001$), and we have no clear interpretation of this fact. Low-volume

¹¹<https://web.archive.org/web/20181201124839/http://statline.cbs.nl/Statweb>.

Table 1 Correlations between hotel visits and GTI in Amsterdam and in the Netherlands (2012–2017). Source: Statistics Netherlands. $N = 294$ for each test. Significance level: (***) $p < 0.001$, (**) $p < 0.01$, (*) $p < 0.05$

Target	Low-volume	GTI levels	Kendall's τ
Amsterdam	No	12	0.57***
–	Yes	11	0.13**
Netherlands	No	32	–0.03
–	Yes	25	–0.13**

data shows weak correlations in both cases. This indicates that low-volume data is noisier and needs extra caution to be handled. Bearing in mind the specificity of the case study, these results suggest that searches for cities correlate fairly well with hotel visits, while no correlation is visible with searches for countries (RQ3). Intuitively, this might be explained with the fact that hotel are searched for in specific cities, and not in whole countries.

4 Spatial Analysis of Interest in Municipalities

After studying the global interest origin for Amsterdam and the Netherlands, we turned our attention to municipalities (*gemeenten* in Dutch) as interest destinations in the Amsterdam metropolitan area, which covers about 2,600 km² and currently hosts 2.3M residents. Its 33 spatial administrative units represent Amsterdam and the cities surrounding it. The largest municipalities in terms of resident population are Amsterdam (about 845,000), Almere (201,000), and Haarlem (159,000), while the smallest municipalities have as few as 6,000 residents.

4.1 Data Granularity and Comparability

Retrieving data about these municipalities from Google Trends involves distinct technical challenges. The GTI is computed relative to a set of up to 5 search targets. Since the index is expressed as a percentage of the maximum amount of searches over all targets, indices from different queries are not comparable (for example, GTI = 10 for Amsterdam would indicate a lot more searches than the same GTI for a less popular place). To overcome this limitation, we included a reference term in each query which was guaranteed to define the maximum for all target objects, namely “Amsterdam”. This effectively made all municipality indices comparable with the reference term, and thus comparable to each other. A second strategy was to compare the temporal behaviour between regions, in terms of decline or increase in percentage.

A second difficulty lies in the fact that the GTI has a very limited resolution. GTI values near zero are rounded to zero, and all other values are rounded to integers. This limitation is enforced by design to avoid exposing fine-grained information that could be used to alter rankings. This makes it difficult to compare index distributions over interest targets that are spread out with a long tail in higher values, because many low values are reduced to zero or 1.

In order to collect data at a sufficient granularity, we implemented an algorithm that increases the resolution of a given query. This is done by subdividing the query terms along their order of magnitude at GTI gap of more than 80, then re-querying the lower subset of terms, and finally scaling the new index values of the subset to the old index value. For example, one can query the GTI for “Amsterdam”, “Almere” and “Beemster” for 2018. The result is Amsterdam = 100 (base term), Almere = 6, and Beemster < 1. Since the GTI gap is more than 80, a second query is executed with Almere as the base term (100), and Beemster obtains 2. We can then estimate the GTI for Beemster as $2/100 \cdot 6 = 0.12$. While this procedure can in principle be repeated for all index gaps, we only applied it once, since the differences between terms of lower interest turned out to be smaller than 80.

For each municipality, we inspected Google Trends and ensured that the query was fairly unambiguous, i.e. the municipality name did not have another obvious meaning. Interestingly, smaller municipalities were under-represented in Google Trends, with undefined GTI in most instances (RQ1). This suggests that usable data exists for countries, cities, and for some popular neighbourhoods and points of interest (e.g. Van Gogh Museum), but smaller areas are not mappable.

Using the aforementioned algorithm, we collected the GTI *between* municipalities, for period 2007–2017, using a base term to make them comparable at a given time (e.g. was the interest higher for Amsterdam or for Almere in 2007?). The data was then collected *within* municipalities, looking at the temporal variation for a single municipality (e.g. was the interest for Almere higher in 2007 or in 2008?). In both datasets, the GTI excluded low-volume data.

4.2 Interest Between Municipalities

In this analysis, we used the aforementioned algorithm to be able to compare the interest in target municipalities across space, starting from “Amsterdam” as a base term (GTI = 100). Note that all the following maps of the Amsterdam metropolitan region are projected with UTM (zone 31N). The boundaries and population data are from dataset *CBS Wijk- en Buurtkaart 2017*.¹² Figure 4 shows the GTI of the 33 municipalities, ranging from 0 to 62.5.

As can be seen, high-interest municipalities emerged in the suburban area (RQ2). On the Western coast, Zandvoort is a popular beach area close to Amsterdam. Nearby,

¹²<https://web.archive.org/web/20181205153226/https://www.cbs.nl/nl-nl/dossier/nederland-naal/geografische%20data/wijk-en-buurtkaart-2017>.

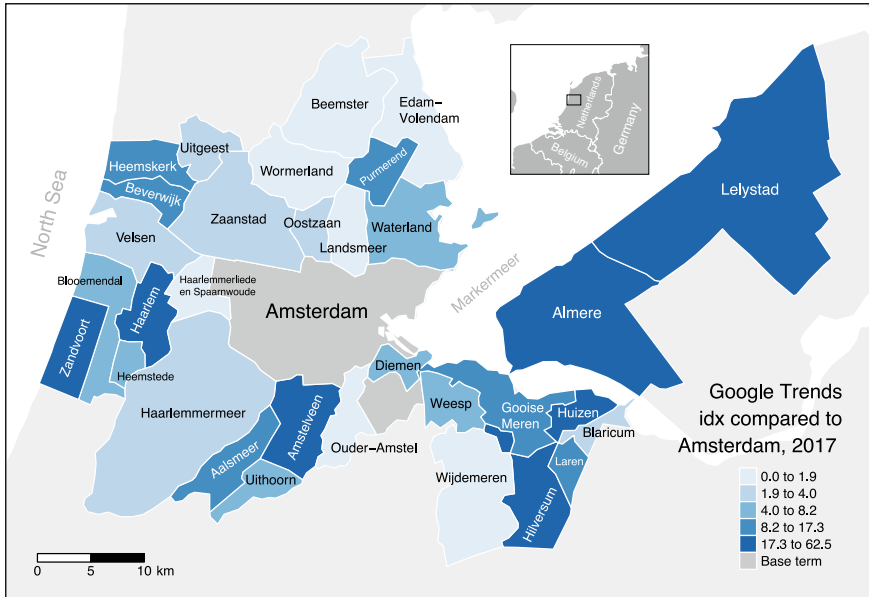


Fig. 4 Google Trends index for municipalities in the Amsterdam metropolitan region in 2017, using “Amsterdam” as the base term, from the whole world, without low-volume data. White areas represent water. Data grouped as quantiles. Source: Google Trends

Haarlem is an important historical city with several popular museums. Amstelveen has an important residential function for people working in Amsterdam, but also contains corporate head offices, including that of KLM, and a prominent art museum. Huizen is a former fishing village. Hilversum is known as the Dutch “media city” with a focus on radio and television broadcasting and a related museum on image and sound. Finally, Almere and Lelystad were both developed in the second half of the 20th century, and are currently important centers for commuters in the Amsterdam area, and also have recreational highlights, such as a large shopping centre with modern architecture, a replica of a cargo ship of the Dutch East India Company, and one of the Factory Outlet Centres in the Netherlands.

As the population size of these municipalities varies from about 6,000 to 845,000, we related the GTI with population densities, by grouping population densities and GTI based on quantiles (RQ4). The bi-variate choropleth map in Fig. 5 shows the geography of search interest in relation to population.¹³ The brown municipalities (high population and high interest) dominate the metropolitan area, and are of particular interest for tourists. The light red areas manage to attract high interest with low population: Huizen, Laren, and Blaricum are very affluent municipalities, and a popular beach is located in Zandvoort. Haarlemmermeer and Velsen, in contrast, are much more affordable housing areas of lower interest.

¹³Map based on R package: <https://github.com/sdesabbata/BivariateTMap>.

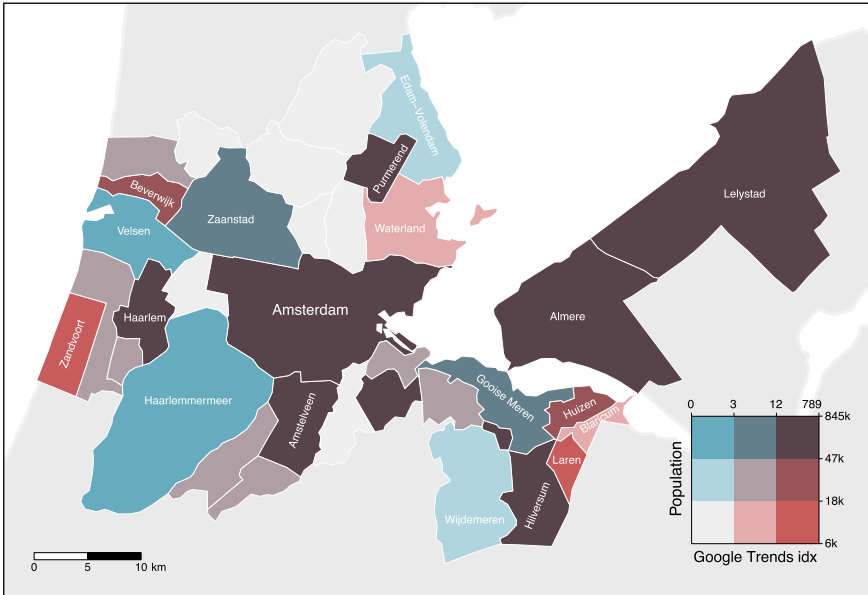


Fig. 5 Population of municipalities in the Amsterdam metropolitan region 2017, in relation to Google Trends index in period 2007–2017 (sum), without low-volume data. The 9 groups are calculated as the intersections of 3 quantiles for the two variables. For the sake of readability, labels are omitted for low-low and medium-medium areas. Sources: Statistics Netherlands and Google Trends

By measuring the *demand* for an area in terms of search interest and its *capacity* in terms of population or infrastructure density, we suggest that the worldwide interest in the red and brown municipalities is likely to overcome their carrying capacity. Thus, rising numbers of residents and visitors can be expected with a further increase in property prices and the development as well as aggravation of capacity problems, including (over)crowding. This Google Trends data allows to identify, among others, Amsterdam, Almere, Lelystad, Haarlem, Hilversum, and Amstelveen as hotpots that combine high population and high tourist in-flows, resulting in overcrowding (RQ4). By contrast, turquoise areas (high population and low interest) emerged as under-represented in the interest geography. These areas might be considered as targets for intervention to absorb tourist flows from high-interest areas (RQ4).

4.3 Interest Within Municipalities

To see further temporal and spatial trends in the Amsterdam metropolitan area (RQ1), we studied the temporal variation in GTI for each municipality separately, comparing only whether interest declined or increased in the last decade. This latter measure is

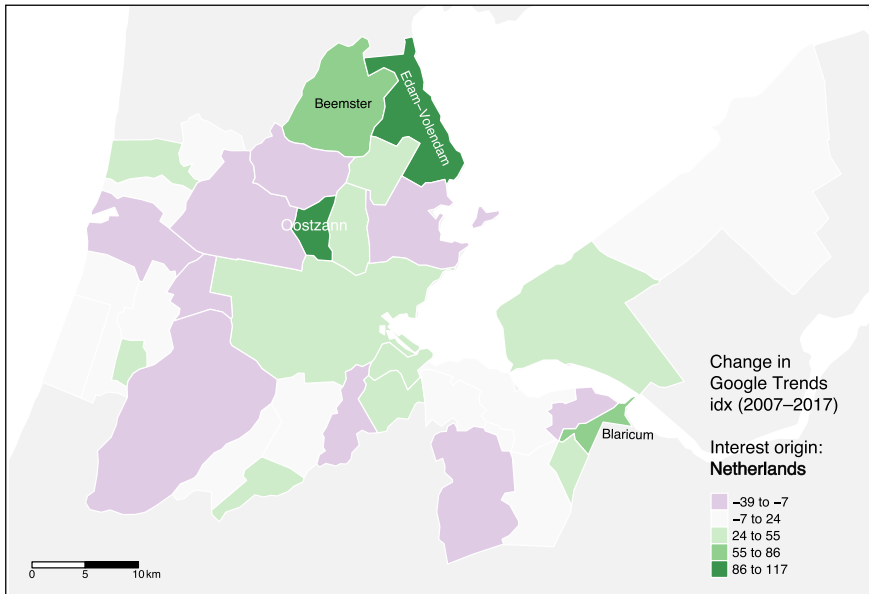


Fig. 6 Change in Google Trends index in each municipality, without low-volume data (2007–2017). Interest origin: Netherlands. Data is grouped with equal intervals. For the sake of readability, some labels are omitted. Source: Google Trends

comparable between areas, while the underlying absolute values are not. Rather than the variation with respect to a base term, in this instance, GTI is 100 at the peak of each municipality.

Figure 6 shows the change in GTI for each municipality, only considering searching originated from the Netherlands. The GTI increased for some municipalities (i.e. Amsterdam, Almere, Oostzaan, and Landsmeer, Edam-Volendam and Beemster, Blaricum and Laren), while other areas declined, including Haarlemmermeer, Zaanstad, and Wormerland. Possible explanations for the interest growth for some areas: Edam-Volendam, an old fishing village with a traditional cheese market, may have become more popular among visitors in recent years. Blaricum and Laren are among the most affluent municipalities in the Netherlands, attracting investors in real estate and luxury shoppers. Almere serves as one of the most important parts of the commuter belt of Amsterdam, and has absorbed some of the housing demand in the area, but also has appealing recreational facilities. Beemster has several historical fortresses that attract tourists. The popularity of Oostzaan may relate to the fact that it is located closely to the Zaanse Schans, a traditional working village with several windmills.

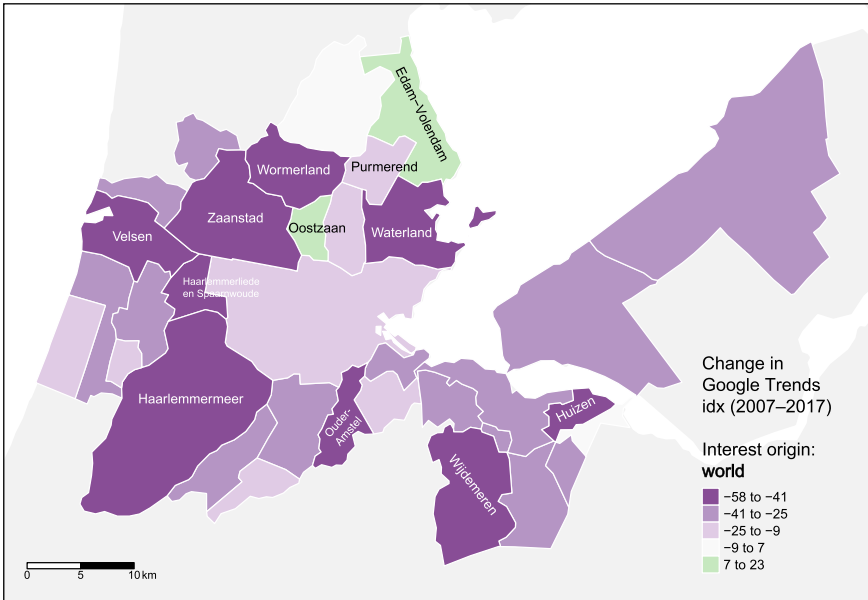


Fig. 7 Change in Google Trends index in each municipality, without low-volume data (2007–2017). Interest origin: world. Data is grouped with equal intervals. For the sake of readability, some labels are omitted. Source: Google Trends

When observing global searches, the distribution of change looks similar, but with a steep decline over most areas (Fig. 7). Even Amsterdam, Almere and Blaricum lost interest over the last decade, while only Beemster, Edam-Volendam and Oostzaan stayed equal or increased. To investigate the decline, it is beneficial to look at the variation in GTI at a higher temporal detail (Fig. 8). Almost all areas declined in 2008–2011, with a particularly steep decline in the Western part of the region, with negligible exceptions. The decline is subsequently more biased towards the Eastern areas, with recovery localized in the North. Finally, from 2014 to 2017, more areas are either recovering or remaining stable.

This overall international decline in interest in the Amsterdam metropolitan area escapes simple explanations, and seems to counter other trends (RQ3). A possible factor can be identified in the sharp increase in mobile searches on smartphones, which broadened and changed the search pool, therefore impacting on the GTI. While the 2008 financial crisis marked a temporary reduction in tourist flows, tourism at the global level seems rather “shock proof” and shows robust growth in the period 2007–2017. This suggests that spatially and temporally more detailed analyses are needed to relate this large decline to ground truth data (RQ2 and RQ3).

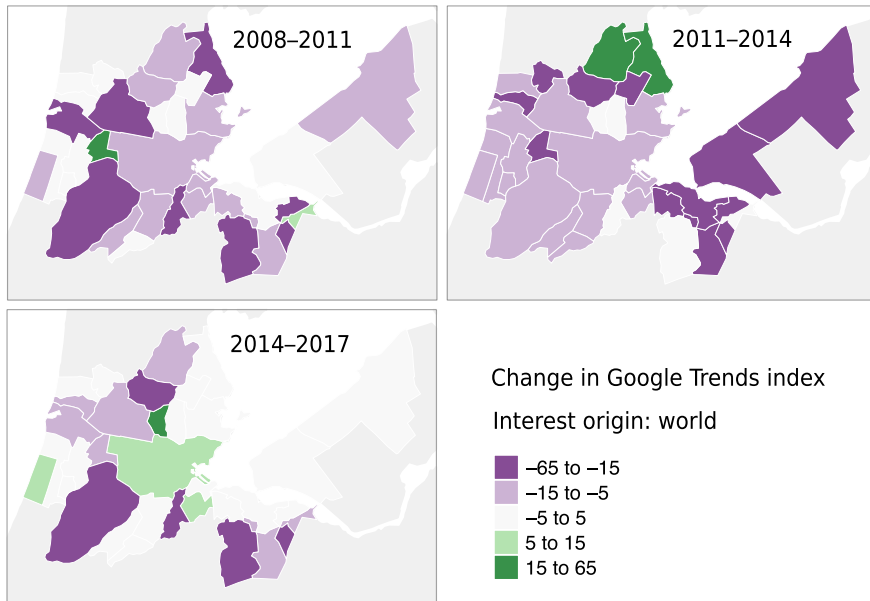


Fig. 8 Change in Google Trends index in each municipality, without low-volume data in 3 temporal intervals (2007–2017). Groups were manually defined. For the sake of readability, some labels are omitted. Interesting origin: world. Source: Google Trends

5 Discussion and Conclusion

In these analyses of Google Trends, we have shown several possibilities for empirical research. First, we studied the interest origin for the Netherlands and Amsterdam at the global level (RQ2), highlighting the rise of Europe as the main origin area, and how post-colonial ties are clearly visible in the spatial distribution of interest. Second, we compared search trends with hotel visits (RQ3). The correlation turned out to be fairly strong at the city level, but very weak at the country level (RQ1).

We then moved on to observe the interest geography at the municipality level, comparing areas to one another and in terms of population (*between* areas) and at the level of individual change (*within* areas). This analysis illustrated how uneven and volatile the geography of interest is, and how some areas appear largely over-represented in the interest they generated (e.g. very affluent municipalities) and others are under-represented (e.g. more affordable residential areas) (RQ4). While this unevenness is perhaps not surprising, it is novel to view interest quantified over time at this scale (RQ1).

For the case study of the Amsterdam metropolitan area, we have shown that it is possible to detect major areas of interest, in particular ones that attract a relatively high volume of searches despite low population density: These are potential hotspots of tourist overcrowding. In addition, we might be able to detect suburban areas

where interest is rising due to tourist attractions and appealing housing opportunities (globally or from a specific country of origin). These are areas where interest, by tourists as well as residents, is already rising. Processes of a rapid increase in interest, such as gentrification, might be detectable through this data. Furthermore, areas that are not yet on the tourist map could be identified for future marketing or development.

Rather than providing exhaustive answers to our research questions, this article opens up a path to further, more modelling-oriented investigations, adopting high-quality, tourism-related datasets as ground truth. While not reaching final conclusions, this exploration has provided us with several insights. In order to make effective use of Google Trends data in tourism studies, and more broadly, in geography, we deem the following points to be noteworthy:

- The geographic scale has a strong impact on the search behaviours for places. For example, patterns at the country and at the city level appear very different (e.g. city searches correlate to hotel bookings, country searches do not). Furthermore, for scales finer than a municipality, the data is not granular enough.
- The resolution of the GTI is intrinsically limited. Our algorithm (see Sect. 4.1) can help increase the resolution of the index resolution for relatively low-interest areas. However, peripheral small areas are likely to have a $GTI = 0$, which cannot be re-scaled. Many prominent points of interest can attract higher interest than large spatial units (e.g. Van Gogh Museum).
- Semantic ambiguity of search terms is a critical problem that cannot be solved completely. The manual inspection of search results for each query in the study is highly recommended. An approach worth pursuing might be to distinguish between tourism, reference, and housing searches through more specific terms (e.g. “Amsterdam hotels” as opposed to “Amsterdam rent”).
- Estimates from SEO companies such as SemRush can provide a quantification of search volume to complement the GTI, although the quality of such data is largely undefined.
- Searches for places tend to exhibit strong seasonality. Interest bursts are a serious issue and should be accounted for. Burst-causing events might be both unplanned (natural and man-made disasters) and planned (e.g. the World Cup).
- It is essential to possess local expert knowledge in order to interpret trends and spatial patterns both in interest origin and targets. Even so, in our case study, some patterns remained hard to interpret with obvious explanations (e.g. tourist flows). Place searches have their own peculiar, rather volatile geography, and some hard-to-explain variation must be expected.

This study of the spatial dimension of Google Trends data can enable a novel observation point on urban geographies. More efforts are needed to devise techniques to increase the semantic accuracy of terms, to reduce noise from bursts, and to disaggregate search behaviours that are currently conflated. Other kinds of ground-truth flow data (tourism and migration statistics, Airbnb stays, and smartphone data) could in the future be used to calibrate the Google Trends statistics. The inclusion of more granular search targets, such as a specific, highly visible points of interest,

may also uncover meaningful patterns. Such big data sources might help gather new knowledge in the geographic domain, including the tourism-induced (over)crowding crisis that places increasing pressure on several cities around the world.

Acknowledgements The authors gratefully acknowledge Google for making some of its search data publicly available, Flavio Ponzio for providing insights on Search Engine Optimisation, and Stefano De Sabbata for his bi-variate choropleth library.

References

- Askitas N, Zimmermann KF (2015) The internet as a data source for advancement in social sciences. *Int J Manpow* 36(1):2–12
- Ballatore A, Graham M, Sen S (2017) Digital hegemonies: the localness of search engine results. *Ann Am Assoc Geogr* 107(5):1194–1215
- Ballatore A, Jokar Arsanjani J (2018) Placing Wikimapia: an exploratory analysis. *Int J Geogr Inf Sci* 1–18
- Ballatore A, Scheider S, Lemmens R (2018) In: Mansourian A, Pilesjö P, Harrie L, van Lammeren R (eds) *Patterns of consumption and connectedness in GIS web sources*. Geospatial technologies for all. Agile 2018. Springer, Berlin, pp 129–148
- Choi H, Varian H (2012) Predicting the present with Google Trends. *Econ Rec* 88(1):2–9
- Dergiades T, Mavragani E, Pan B (2018) Google trends and tourists' arrivals: emerging biases and proposed corrections. *Tour Manag* 66:108–120
- Goel S, Hofman JM, Lahaie S, Pennock DM, Watts DJ (2010) Predicting consumer behavior with web search. *Proc Natl Acad Sci USA* 107(41):17486–17490
- Gospodini A (2006) Portraying, classifying and understanding the emerging landscapes in the post-industrial city. *Cities* 23(5):311–330
- Graham M, De Sabbata S, Zook MA (2015) Towards a study of information geographies: (im) mutable augmentations and a mapping of the geographies of information. *Geo Geogr Environ* 2(1):88–105
- Hall T, Barrett H (2017) *Urban geography*, 5th edn. Routledge, London
- Hendler J, Shadbolt N, Hall W, Berners-Lee T, Weitzner D (2008) Web science: an interdisciplinary approach to understanding the web. *Commun ACM* 51(7):60–69
- Jun S-P, Yoo HS, Choi S (2018) Ten years of research change using Google Trends: from the perspective of big data utilizations and applications. *Technol Forecast Soc Chang* 130:69–87
- Kitchin R (2014) Big data, new epistemologies and paradigm shifts. *Big Data Soc* 1(1):1–12
- Lazer D, Kennedy R, King G, Vespignani A (2014) The parable of Google Flu: traps in big data analysis. *Science* 343(6176):1203–1205
- Li X, Pan B, Law R, Huang X (2017) Forecasting tourism demand with composite search index. *Tour Manag* 59:57–66
- Neuts B, Vanneste D (2018) Contextual effects on crowding perception: an analysis of Antwerp and Amsterdam. *J Econ Soc Geogr* 109(3):402–419
- Novy J, Colomb C (2017) *Protest and resistance in the tourist city*. Routledge, London
- Önder I, Gunter U (2016) Forecasting tourism demand with Google Trends for a major European city destination. *Tour Anal* 21(2–3):203–220
- Pan B, Chenguang Wu D, Song H (2012) Forecasting hotel room demand using search engine data. *J Hosp Tour Technol* 3(3):196–210
- Popp M (2012) Positive and negative urban tourist crowding: Florence, Italy. *Tour Geogr* 14(1):50–72

- Silverstovs B, Wochner DS (2018) Google trends and reality: Do the proportions match?: appraising the informational value of online search behavior: evidence from swiss tourism regions. *J Econ Behav Organ* 145:1–23
- Singleton AD, Spielman S, Brunsdon C (2016) Establishing a framework for open geographic information science. *Int J Geogr Inf Sci* 30(8):1507–1521
- Spierings B (2013) Fixing missing links in shopping routes: reflections on intra-urban borders and city centre redevelopment in Nijmegen, The Netherlands. *Cities* 34:44–51
- Stephens-Davidowitz SI (2013) *Essays using Google Data*. PhD thesis, Harvard University, Cambridge, MA
- Yang X, Pan B, Evans JA, Bendu Lv (2015) Forecasting Chinese tourist volume with search engine data. *Tour Manag* 46:386–397

Estimating the Spatial Distribution of Vacant Houses Using Public Municipal Data



Yuki Akiyama, Akihiro Ueda, Kenta Ouchi, Natsuki Ito, Yoshiya Ono, Hideo Takaoka and Kohta Hisadomi

Abstract This study aimed to develop a new method for estimating the spatial distribution of vacant houses using municipal public big data and sample field surveys instead of a field survey of the whole municipal area. This can help reduce the cost, time, and labor involved in conducting vacant house surveys. For this purpose, we developed a vacant house database to integrate various public big data with field survey results for different parts of Japanese municipalities (Kagoshima and Asakura). With our newly developed method, we could estimate the spatial distribution of vacant houses with high reliability utilizing crosstab tables developed from the database. The results help to realize a method for conducting surveys of vacant house distribution in broad areas in a quick, inexpensive, and continuous manner, which has not yet been achieved by previous studies.

Keywords Vacant house · Estimation · Public municipal data · Spatial distribution

Y. Akiyama (✉)

Center for Spatial Information Science, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa-shi, Chiba, Japan

e-mail: aki@csis.u-tokyo.ac.jp

A. Ueda · K. Ouchi · N. Ito · Y. Ono

MLIT, 2-1-2, Kasumigaseki, Chiyoda-ku, Tokyo, Japan

e-mail: ueda-a87ck@mlit.go.jp

K. Ouchi

e-mail: ohuchi-k2mw@mlit.go.jp

N. Ito

e-mail: itou-n2uw@mlit.go.jp

Y. Ono

e-mail: oono-y2cm@mlit.go.jp

H. Takaoka

Japan Real Estate Institute, 1-2-3, Kaigan, Minato-ku, Tokyo, Japan

e-mail: hideo-takaoka@jrei.jp

K. Hisadomi

Zenrin Co. Ltd, 101, 2-Chome, Awajicho, Chiyoda-ku, Tokyo, Japan

e-mail: kotayo2588@zenrin.co.jp

© Springer Nature Switzerland AG 2020

P. Kyriakidis et al. (eds.), *Geospatial Technologies for Local and Regional Development*, Lecture Notes in Geoinformation and Cartography, https://doi.org/10.1007/978-3-030-14745-7_10

1 Introduction

In recent years, the number of vacant homes in Japan has continued to increase due to the declining population, the aging society, and migration to large cities. According to the Housing and Land Survey conducted in 2013 by the Statistics Bureau of Japan, the number of vacant houses in 2013 was 8.20 million, and the rate of vacant houses was 13.5%, showing continuous increases since the 2003 survey results. In particular, among vacant houses, the increase in the number of houses called “other houses” where no one lives currently is remarkable. The number of such houses was 2.12 million nationwide in 2003, increasing to about 3.18 million in 2013, according to the Housing and Land Survey. It is estimated that this number will continue to increase in the future. For example, Akiyama and Shibasaki (2015) estimated that the number of other vacant houses will increase from 3.18 million to 4.17 million from 2013 to 2040. Moreover, Imai et al. (2015) estimated that the number of such houses will reach about 21.5 million in 2033 in the pessimistic scenario.

Many previous studies have suggested that an increase in vacant houses in Japan can increase crime, adversely affect the landscape, and reduce the charm and vitality of surrounding areas (Abe et al. 2014; Asami 2014; Baba and Asami 2017). In the US context, Accordino and Johnson (2002) found that an increase in vacant houses and lots in certain cities greatly affected neighborhood vitality, crime prevention efforts, and commercial vitality. Similar effects have been identified in other US cities (Zahirovich-Herbert and Gibler 2014) as well as in South Korea (Nam et al. 2016). Meanwhile, different from Japan, Turnbull and Zahirovic-Herbert (2010) found that “vacant houses enjoy stronger shopping externality effects from surrounding houses for sale than do their occupied counterparts.” In Japan, because of the accumulation of damage from severe climatic conditions (e.g., typhoons, heavy snowfall, and earthquakes) and frequent changes to the Building Standards Act, Japanese people tend to avoid moving into secondhand houses. Therefore, once a house becomes vacant, especially an old one, there is a high likelihood that new residents will not move in.

In response, the “Act on Special Measures to Forward Municipalities’ Moves for Vacant Premises” was put into effect in 2015. Under this act, Japanese municipalities nationwide are promoting measures to cope with vacant houses. Per this act, to monitor the spatial distribution of vacant houses in municipalities, all Japanese local governments are required to develop databases of vacant houses. Therefore, there is an urgent need for a method that can quickly, easily, and inexpensively monitor or estimate the spatial distribution of vacant houses in a municipality using only data owned by the municipality.

The main method for monitoring the spatial distribution of vacant houses in a broad area has been to visit individual buildings and judge them by looking at their exteriors. Even today, methods are not well established to estimate areas where many vacant houses are distributed and thus prioritize field-survey areas, to monitor or estimate the spatial distribution of vacant houses in a broad area, and to periodically update information on vacant house distribution. Therefore, surveying vacant houses

costs local governments a lot of time, money, and labor. For example, in a 2017 study by the authors, it took about half a year and 10 million JPY to survey an area with about 4000 vacant houses in a city in Osaka Prefecture. It would therefore be very difficult for municipalities with weak fiscal capacity to conduct such surveys on a regular basis. This poses a major barrier to accelerating measures to deal with vacant houses in Japan.

2 Related Work

Several studies have attempted to monitor and estimate the spatial distribution of vacant houses in Japan. For example, Ishikawa and Akagawa (2014) and Kubo and Mashita (2015) tried to clarify the occurrence factor of vacant houses. Kato et al. (2009) investigated the current state of management methods for vacant houses. Sasaki et al. (2010) investigated the utilization situation of vacant houses. Many of these studies analyzed vacant houses from various perspectives after monitoring their spatial distribution. However, in many such studies, the method for monitoring the distribution of vacant houses was based on field survey by visual inspection or interviews with local residents or residents' associations. It is difficult to apply such methods for continuous investigations in broad areas, such as entire municipalities.

Therefore, in recent years, some studies have tried to monitor the spatial distribution of vacant houses by utilizing various statistics and spatial information. For example, Yamashita and Morimoto (2015) estimated the distribution of vacant houses in a city based on water closure information; however, this method defined all buildings where water supply had ceased as vacant houses, and the rationale for that definition was unclear. Ishikawa et al. (2017) estimated the vacant house rate of each city block using point data for building locations and the Japanese population census; however, the reliability of the method was not verified in comparison with field-survey results. Outside of Japan, Accordino and Johnson (2002), Nam et al. (2016), and others have attempted to estimate the number of vacant homes. However, they considered macro aggregation units, such as municipal units, using existing national or municipal statistics and were thus unable to understand the distribution of vacant houses in small areas.

There are examples of Japanese municipalities estimating vacant-house distribution using water closure information in Kyoto in 2014, Nara in 2015, and so on. However, the survey areas were very limited, and there have been no attempts so far to monitor or estimate the distribution of vacant houses in entire municipal areas by methods other than field surveys. Additionally, data on vacant house surveys developed by private companies have also recently become available in Japan. However, since there are charges for these data, it is difficult for many Japanese municipalities facing declining revenues (due to population decline) to acquire and continuously update the data.

3 Study Objective

In light of the abovementioned problems, this study aimed to develop a method for local governments to quickly, easily, and inexpensively conduct vacant house distribution surveys. This is not a conventional field survey method that visually inspects the whole target area. Rather, by using field-survey results from some sample areas and various municipal big data (called “public big data” in this paper) owned by municipalities, we developed a method to quickly estimate the spatial distribution of vacant houses in a broad area—namely, the whole municipal area. The reason for mainly using public big data is that it is desirable to conduct surveys at a low cost using data already possessed by municipalities as much as possible.

This study also aimed to understand the distribution of vacant houses among detached houses. This is because there are differences in many aspects, including the method of monitoring spatial distribution, between detached houses and apartment houses, as well as for promoting the utilization of vacant houses and understanding the distribution of vacant houses that are likely to become “other houses,” which are expected to increase in the future. Moreover, this study aimed not to accurately specify the distribution of each vacant house but to estimate and understand the areas (city blocks or grid units) where many vacant houses are distributed.

Our method has great usability in municipalities. First, the method proposed in this study can solve the problems encountered by previous studies—in particular, the difficulty of rapidly investigating the spatial distribution of vacant houses in a wide area. Second, our method was realized using public big data possessed by municipalities. Public data are updated continuously every year as a part of regular municipal tasks, so there is little additional financial cost for municipalities. In other words, the method can contribute to realizing inexpensive and continuous surveys. Third, our method ensures versatility because it uses public big data that are commonly maintained by municipalities throughout Japan. Finally, the method is highly applicable outside Japan since similar data are being developed in municipalities in other countries. In this way, our method can help overcome the limitations of previous studies.

4 Field Survey

First, to understand the spatial distribution of vacant houses from public big data, it is necessary to analyze the characteristics of a vacant house. For this purpose, information about a certain number of actual vacant houses is needed. Therefore, we first extracted sample areas from the target areas of this study and then collected the actual locations of vacant houses by field survey through visual inspection of each house.

The study areas were Kagoshima city in Kagoshima Prefecture and Asakura city in Fukuoka Prefecture. These were selected because these municipalities allowed us

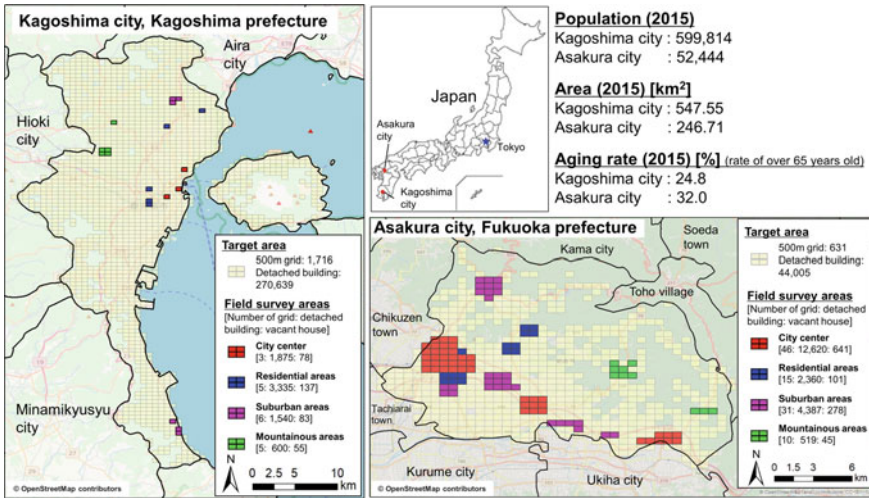


Fig. 1 Study areas and field survey areas

Table 1 Criteria for definite vacant houses and non-vacant houses

Criteria for a non-vacant house	<ul style="list-style-type: none"> – Electricity meters are working – People come in and out of the house – Laundry is dried – Flower beds are kept – There is a light on in the house – There are cars that are used now, etc.
Criteria for a definite vacant house	<ul style="list-style-type: none"> – Electricity meters are not working – There are no traces of people coming in and out of the entrance – There is no furniture in the house – The mailbox is closed – There are signs for sale or rent – The house has collapsed and people cannot dwell in it

to conduct field surveys in a relatively wide area, and they cooperated in providing public big data. Figure 1 shows the study areas and sample areas of the field survey. These included an appropriate balance of residential areas, the city center, suburban areas, and mountainous areas.

In the field survey, the standards presented in Table 1 were set, and all detached houses in the field survey areas were classified into the following three categories according to the procedure shown in Fig. 2. Buildings that met at least one non-vacant house criterion were classified as “non-vacant houses”; those that met at least one definite vacant house criterion were categorized as “definite vacant houses”; and those that did not meet any criterion were “estimated vacant houses.” In this study, we decided to treat the definite vacant houses as having the real value of a vacant

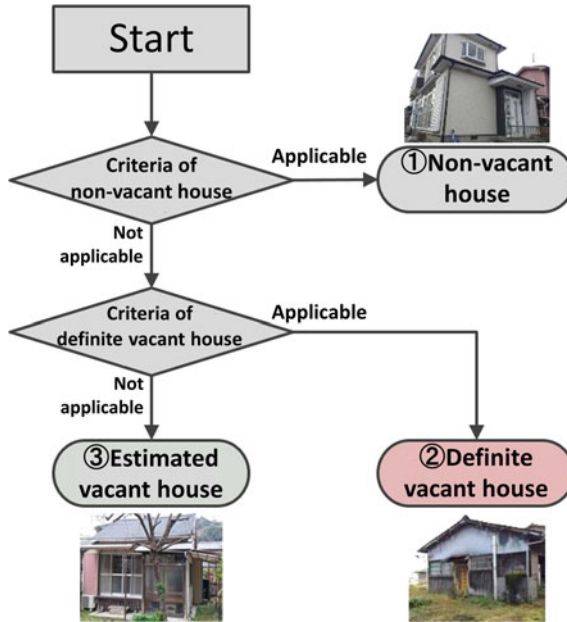


Fig. 2 Evaluation process for vacant houses in the field survey



Fig. 3 Photographs of field surveys and definite vacant houses in Kagoshima

house. Figure 3 shows a photograph of a field survey and definite vacant house in Kagoshima. Table 2 shows the results of the field surveys.

Table 2 Date of field survey and number of surveyed detached buildings and vacant houses

City name	Date of field survey	Number of surveyed detached buildings	Number of surveyed vacant houses	Rate of vacant houses (%)
Kagoshima	August 22–26, 2016	7350	353	4.80
Asakura	October 17–21, 2016	19,886	1065	5.36

5 Development of Vacant House Database

We developed a database for analyzing the characteristics of vacant houses by spatial integration of the real value of vacant houses collected by the field survey and municipal big data obtained from municipalities with detached building polygon data recorded on the digital residential map across the study area using GIS. We call this database as the “Vacant house database” in this study. Table 3 shows the attribute list for the vacant house database. All buildings in the database contain six kinds of data sources, as shown in Table 3.

“(1) Residential map (RM),” in Table 3, provided digital data (polygon data) on each building with attributes such as building name, use, area, and number of floors of all buildings in Japan. These are commercial data; however, since they have already been introduced by many municipalities in Japan and the barriers to obtaining them are low, this study treats them as data owned by municipalities. “(3) Basic resident register (BRR)” provided digital data of resident cards and official information on all residents living in each municipality. “(4) Hydrant consumption amount information (HCI)” is the database on water usage by houses managed by the municipal waterworks bureau. (5) Building registration information (BRI) is information on individual land and buildings managed by municipalities and the Legal Affairs Bureau. “(6) National land numerical information” is various data on Japan’s land area developed by the Ministry of Land, Infrastructure, Transport and Tourism. It is open data.

BRR, HCI, and BRI are public big data, and all public big data are not open data, although they can be used within the municipality. Therefore, to carry out this study, we use them after approval for research use by the Personal Information Protection Review Board of Kagoshima and Asakura, and we cannot collect personal names due to anonymized processing.

However, as shown in Table 4, public big data were not necessarily given to all detached buildings. The reason is that those buildings whose names are registered as unknown on the residential map contain not only vacant houses but also an attached building such as a warehouse or a garage. Since these buildings are not used for residential purposes, many buildings were not integrated with the public data. In addition, the reason why the integration rate of the HCI in Asakura is low is that

Table 3 Attributes of the vacant house database

Data source	Survey date	Number of records ^b	Data size (MB) ^b	Attribute	Data type
(1) Residential map (RM)	K: 2016 A: 2016	K: 285,917 A: 44,692	K: 667.4 A: 108.1	Building ID	V
				Building use	C
				Building area (m ²)	V
(2) Field survey	K: Aug. 22–26, 2016 A: Oct. 17–21, 2016	–	–	Field survey area name (in case of field survey area)	C
				Evaluation result of vacant house	C
(3) Basic resident register (BRR)	K: Jun. 2016 A: Sep. 2016	K: 292,285 A: 120,326	K: 82.5 A: 8.9	BRR could be integrated (or not) with RM ^a	C
				Number of households	V
				Number of residents	V
				Average age of residents	V
				Youngest resident	V
				Oldest resident	V
				Shortest length of residence (years)	V
Longest length of residence (years)	V				
(4) Hydrant consumption amount information (HCI)	K: Jun. 2016 A: Oct. 2016	K: 270,218 A: 167,530	K: 16.1 A: 2.9	HCI could be integrated (or not) with RM (Non-integrated, Integrated and opened, Integrated and closed) ^a	C
				Duration of hydrant closure from last closed year (year)	V
				Annual consumption amount (t)	V
(5) Building registration information (BRI)	K: Jan. 2016 A: Jan. 2016	K: 253,511 A: 33,804	K: 16.6 A: 1.6	BRI could be integrated (or not) with RM ^a	C
				Building age (year)	V
				Building structure	C
				Building use	C
(6) National land numerical information	K: 2011 A: 2011	–	–	Designated land use zone	C

Note “V” in the “data type” column means numerical data, and “C” means categorical data. “K” in the “survey date” column means Kagoshima, and “A” means Asakura

^aBRR, HCI, and BRI could not be integrated with RM because, while RM contained all buildings, it did not cover all buildings. Therefore, there was no information for some buildings in RM

^bNumber of records and data size of the source data (residential maps in image file formats and public data in text file formats)

Table 4 Spatial integration rate of public data with building polygon data of residential map

City	Kind of bldg.	Num. of bldg.	BRR		HCI		BRI	
			NSI ^a	Rate [%]	NSI ^a	Rate (%)	NSI ^a	Rate (%)
Kagoshima	All bldg.	7350	3708	50.45	3793	51.61	4180	56.87
	Bldg. with name	4709	3558	75.56	3627	77.02	3569	75.79
	Vacant house	353	70	19.78	79	22.25	194	55.01
Asakura	All bldg.	19,886	7278	36.60	4474	22.50	7465	37.54
	Bldg. with name	8857	6250	70.57	3981	44.95	6195	69.94
	Vacant house	1065	235	22.09	108	10.17	388	36.40

^a“NSI” means the number of spatial integration rate of public data

some houses in Asakura, especially in the suburbs and mountainous areas, still use well water. Although, spatial integration rates of public data with vacant house were smaller than all building and buildings with name. It means that BRR, HCI, and BRI expect to estimate detached building is vacant or not.

6 Development of Estimation Method of Vacant House Distribution

6.1 Overview of the Method

The field survey data included in the vacant house database have the attributes shown in Table 3 for each building. Therefore, this study performed a cross tabulation using these attributes as explanatory variables, and calculated the vacancy rate for each combination. By allocating the result to each building, we estimated the vacancy rate of each building. In this study, we call each combination of crosstabs a “cell.”

First, this study determined the number of used explanatory variables for the estimation of vacant house rate to create a cross tabulation table using all field survey data collected from Kagoshima and Asakura and compare the true value of the vacant house and the estimated value every 250 m grids by correlation analysis for each number of used variables. In addition, we need to verify how reliably the distribution in the entire area can be estimated using the data of the field survey area; the estimation of the vacant house distribution of the whole city based on the field survey results is shown in Fig. 1. Therefore, we developed a cross tabulation table by randomly using 50% data obtained from the field survey areas and verified the reliability of the method by comparing the estimation results with the true value using the same cross tabulation table in both cities. Moreover, the same random sampling was repeated 1000 times, and 1000 samples of cross tabulation were created using all samples.

Finally, we verified the variation of reliability by comparing the estimation results and the true value using these crosstab tables in both cities.

Incidentally, logistic regression analysis is known to stochastically determine the state of a response variable (dependent variable) using multiple explanatory variables (independent variables). However, not all explanatory variables used in this paper were integrated with detached buildings, as shown in Table 3. Therefore, it is also necessary to consider cases where each explanatory variable cannot be used. Considering that it is widely used for vacant house distribution surveys in municipalities, we should develop an estimation method that can be used even when certain explanatory variables cannot be employed, as in the method proposed in this paper. Therefore, we did not adopt the logistic regression analysis.

6.2 *Cell Splitting of Explanatory Variables*

First, we performed cell splitting of the explanatory variables in Table 3. The qualitative variables—two kinds of building use, building structure, and designated land use zone—do not require new cell division. However, for the other quantitative variables, there is a need to determine thresholds for cell splitting. For example, Fig. 4 shows the vacancy rate for each number of residents obtained from BRR in all field survey areas. The vacant house rate of buildings with one resident is considerably larger than that of buildings with two or more residents. Additionally, an increase in the number of residents substantially reduces both the number of buildings and the vacancy rate; the rate of buildings with six or more residents is almost zero. From Fig. 4, it can be assumed that it is appropriate to divide the cell into six cells with one to five residents and six or more residents. Thus, it is possible to compress the calculation amount by combining the cells with few samples or similar vacancy rates. The thresholds of other quantitative variables were also determined using the same method. Table 5 shows a threshold list of explanatory variables, using the results obtained from the vacant house database.

6.3 *Development of Crosstab Tables*

Second, by combining 15 types of explanatory variables shown in Table 5, we developed crosstab tables and compared them with the true value of vacant houses to select the crosstab table that matches the true value the most. All crosstab tables were developed using all field survey data in both cities. Moreover, the comparison of the estimated result with the true value of vacant houses is based on the correlation analysis between the true number and the estimated number of vacant houses, estimated by crosstab tables accumulated by every 250 m square grid.

If the number of variables to be used is less than 15, a method of performing calculation on all possible combinations is also conceivable. However, it is realistic

Table 5 Threshold list of variables for cell splitting

V. no.	Variables	Thresholds	Num. of splitting
<i>Qualitative variables</i>			
1	Building use of residential map	“Detached house,” “Detached office,” “Landmark bldg.,” “Detached bldg. without bldg. name,” and “Others”	5
2	Building use of BRI	“Residential bldg.,” “Non-residential bldg.,” “Unknown,” and “N”	4
3	Building structure of BRI	“Wooden,” “Non-wooden,” “RC/SRC,” “Unknown,” and “N”	5
4	Designated land use zone	13 kinds of zones and non-designated	14
<i>Quantitative variables</i>			
Residential map			
5	Building area (m ²)	25, 50, 75, 100, 150, 200, 300, 500, 750, over 1000	12
Basic resident register (BRR)			
6	Number of households	1, 2, 3, over 4, and “N”	5
7	Number of residents	1, 2, 3, 4, 5, over 6, and “N”	7
8	Average age of residents	20, 40, 50, 60, 70, 80, 90, over 100, and “N”	10
9	Youngest resident	10, 30, 40, 50, 80, 90, over 100, and “N”	9
10	Oldest resident	20, 30, 60, 70, 80, 90, over 100, and “N”	9
11	Shortest years of residence of households	5, 10, 20, 30, 40, 50, 70, 90, over 100, and “N”	11
12	Longest years of residence of households	30, 50, 60, 70, 80, over 90, and “N”	8
Hydrant consumption amount information (HCI)			
13	Duration of hydrant closure from last closed year	0, 1, 2, 3, 4, over 5, and “N”	7
14	Annual consumption amount (t)	5, 10, 25, 50, 75, 100, over 150, “N”, and “C (closed)”	10
Building registration information (BRI)			
15	Building age (year)	10, 20, 30, 40, 50, 60, 70, 80, 90, over 100, and “N”	12

Note “V. No.” means variable number, and “N” means no information because public data were not integrated with the residential map

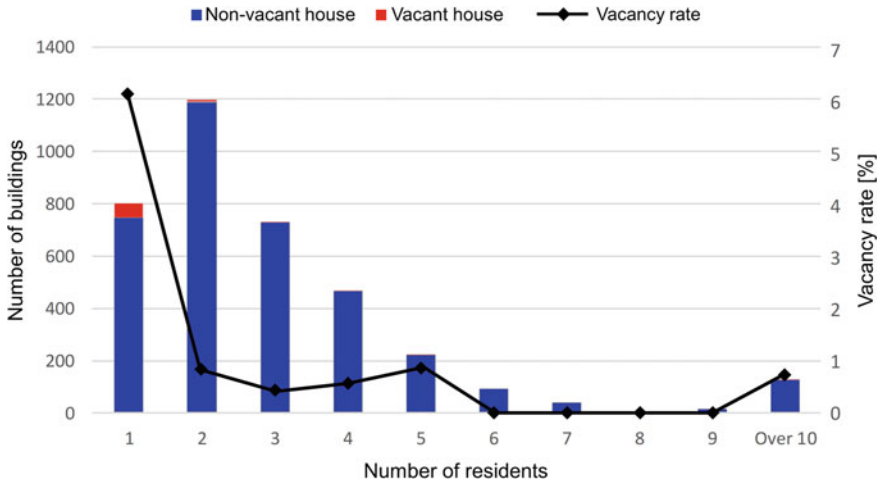


Fig. 4 Vacancy rates for the number of residents obtained from BRR in all field survey areas

to select used variables in advance because this method needs a large amount of calculation. Therefore, we adopted a method of decreasing variables one at a time by repeating the principal component analysis. First, the qualitative variables in Table 5 were categorized by dummy variables. Second, the principal component analysis was performed using all variables; we excluded the smallest variable among the maximum values of the absolute values of each variable included in the principal component whose cumulative contribution rate exceeded 80%. Subsequently, by performing a principal component analysis again, variables with little influence on the principal components were excluded in order. In case of qualitative variables, corresponding qualitative variables were excluded at the stage when all dummy variables related to the variable were eliminated.

6.4 Deciding on the Number of Variables Used

We first estimated the vacancy rate of each building in all field survey areas using crosstab tables. Subsequently, the estimated number of vacant houses for every 250 m square grid in the entire field survey area was estimated by Eq. 1. The 250 m square grid divides the grid in the field survey area of Fig. 1 into two parts equal in length and breadth.

$$\hat{V}_i = \sum_{k=1}^{n_i} \hat{r}_k, \tag{1}$$

Table 6 Reliability of the estimated number of vacant houses accumulated by 250 m square grid using crosstabs by each number of used variables

No. of used variables	Excluded variable no.	Number of cells		Result of correlation analysis	
		Total	Included vacant houses	Correlation coefficient	RMSE
15	–	11,871	696	0.8956	2.4548
14	6	11,824	655	0.8761	2.4551
13	7	11,266	601	0.8552	2.5002
12	5	8786	513	0.8377	2.6711
11	2	8619	491	0.8356	2.6822
10	11	6673	477	0.8327	2.6978
9	13	6562	461	0.8318	2.7026
8	12	5593	439	0.8284	2.7326
7	9	4638	434	0.8246	2.7542
6	10	3185	415	0.8208	2.7772
5	8	1114	263	0.8078	2.8928
4	4	294	126	0.7850	2.9558
3	14	41	33	0.7596	2.9544
2	3	12	12	0.7630	2.9922
1	1	12	12	0.7630	2.9922

Note Numbers in “excluded variable number” match the numbers of “V. no.” in Table 5
p-values in all correlation analyses satisfy <0.0001

where \hat{r}_k is the estimated vacancy rate of detached house k , n_i is the number of detached houses in grid i , and \hat{V}_i is the estimated number of vacant houses in grid i .

Reliability was verified by performing correlation analysis using the estimated number of vacant houses and the true value of the number of vacant houses in each 250 m square grid estimated using Eq. 1. Table 6 shows the reliability using crosstab tables for each variable used in the entire field survey area. The results show that the estimation obtained by the crosstab table created using all 15 variables had the highest reliability. Therefore, from this point forward, we estimated the vacancy rate by using a crosstab table using all variables. The estimation result of the number of vacant houses by the crosstab table using all variables revealed that the correlation coefficient was 0.8956 in units of 250 m square grids, and a very strong correlation was found between the true number and the estimated number of vacant houses.

6.5 Reliability Verification of the Estimation Result Using Random Sampling of Field Survey Results

First, we extracted 50% data randomly from the field survey results, developed a crosstab table using the extracted sample, and compared the estimation result with the true value using the whole field survey area and both cities. Reliability verification was carried out by a correlation analysis of the true value and the estimated value of 250 m square grids, in a manner similar to the previous section. Table 7 shows the results of the reliability verification. The correlation coefficient of the whole field survey area was 0.8598, which was almost the same as the correlation coefficient of the results of using the complete field survey results, 0.8956. Additionally, the results of only Kagoshima and Asakura have almost the same correlation coefficient, and the mean absolute error (MAE) was also low.

Second, the same random sampling was repeated 1000 times, and the results were compared with the true values. Figure 5 shows all correlation coefficients, and Fig. 6 shows all MAEs. The correlation coefficient was about 0.84 on average in any area. Moreover, there was no significant difference in the correlation coefficient between samples. On the other hand, the average MAE was 0.81 in the whole area, 2.09 in Kagoshima, and 0.57 in Asakura, and there was no significant difference in MAE between samples. We estimated the distribution of vacant houses in the whole city area by using the field survey results as sample data, and the result showed the adequacy of estimating the distribution in the wide area by random sampling results.

However, the fluctuation ranges of the correlation coefficient and MAE of Kagoshima were larger than the other results. This is because Kagoshima is a diversified region, which includes a city center of a relatively large-scale local city as well as mountainous villages. On the other hand, Asakura is a small-scale city in the provinces, and the diversity of regional characteristics is small in this city. This problem could be improved by obtaining field survey data on various regional characteristics of other cities and further improving the vacant house database in the future.

Table 7 Result for reliability verification using the 50% sample crosstab table

	Areas		
	Whole area	Kagoshima	Asakura
Correlation coefficient	0.8598	0.8520	0.8583
Adjusted coefficient of determination	0.7377	0.7259	0.7332
MAE	0.6433	1.7876	0.4310
<i>p</i> -value	4.25e-135	2.23e-26	1.99e-120

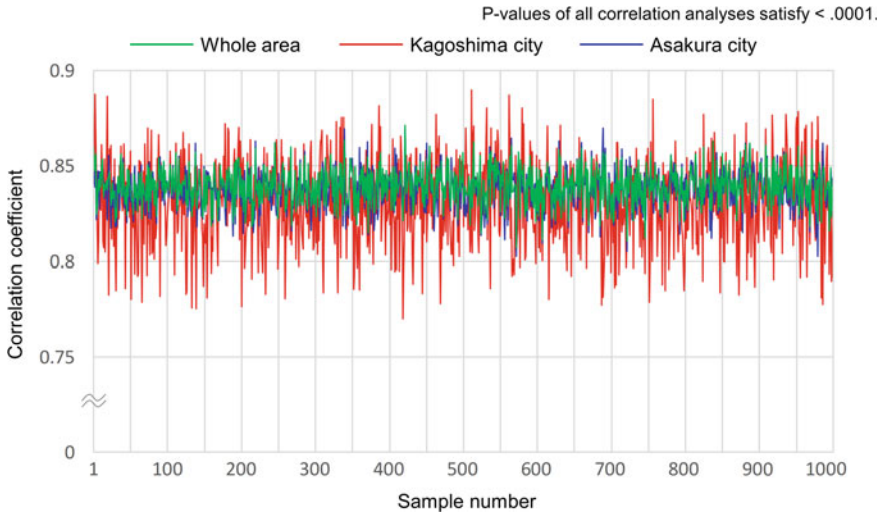


Fig. 5 All correlation coefficients for 1000 random samplings in the whole area, Kagoshima, and Asakura

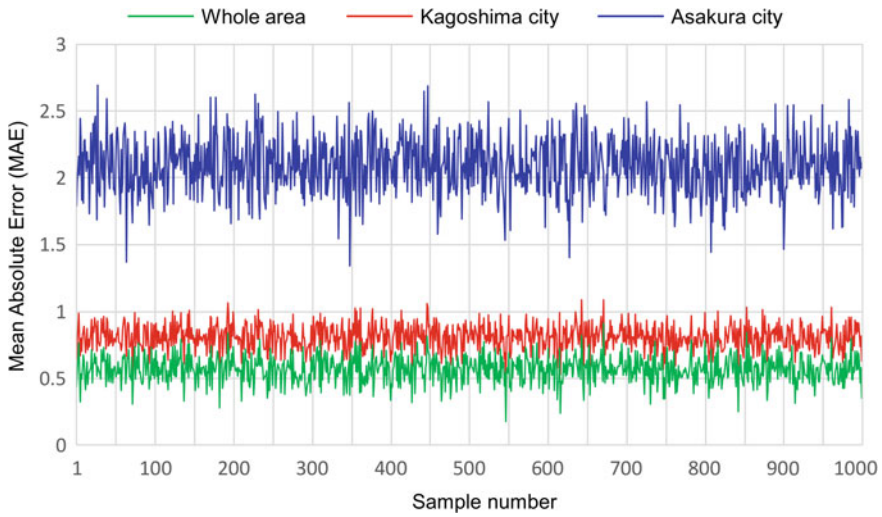


Fig. 6 All MAEs for 1000 random samplings in the whole area, Kagoshima, and Asakura

7 Results

Figures 7 and 8 show the estimated number and rate of vacant houses in 500 m square grid units in Kagoshima and Asakura, respectively, using the crosstabs with the highest correlation coefficients for the whole area in Fig. 5. Figure 7 shows that in Kagoshima, the number of vacant houses is particularly large in residential areas around the central urban area (“A”), residential areas on the plateau (around “B”), old city centers (“C” and “D”), and Sakurajima district affected by active volcano (“E”). Additionally, the vacant house rate is particularly high in the mountainous areas around the city center (around “F”). On the other hand, Asakura municipality was established in 2006 when several municipalities were merged. Figure 8 shows that the number of vacant homes was large in the old central urban area before the merger (“G” and “H”). Moreover, it shows that the number of vacant homes is also large in the old urban area, the castle town (around “I”). The vacancy rate is uniformly distributed in the plain of the south part of the municipality, and the difference between the regions is not as clear as in Kagoshima. Furthermore, the mountainous areas are spread widely in the northern part of the municipality, and the grids are dotted with high vacancy rates in these regions.

Finally, our method could estimate the spatial distribution of vacant houses using public big data in the whole area of Kagoshima and Asakura. However, it is noteworthy

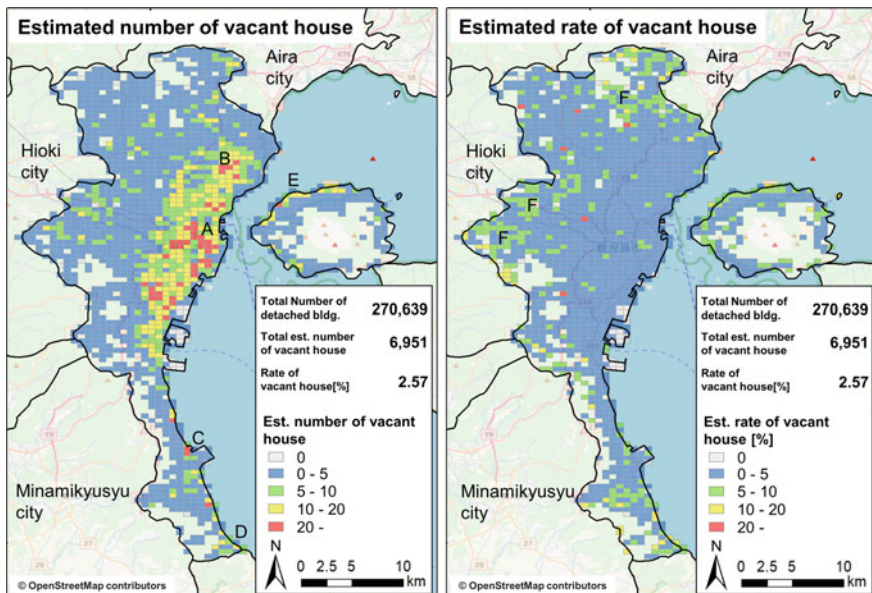


Fig. 7 Estimated number and rate of vacant houses in the entire area of Kagoshima accumulated by 500-m square grids

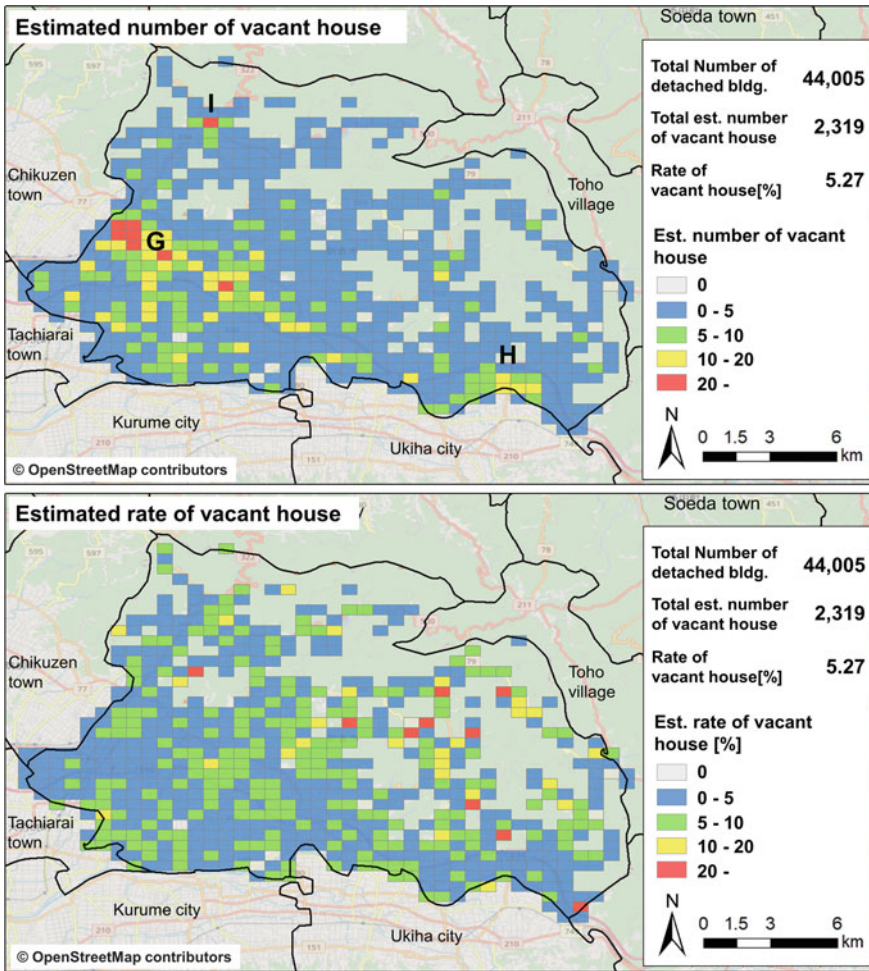


Fig. 8 Estimated number and rate of vacant houses in the entire area of Asakura accumulated by 500-m square grids

thy that our results are an estimation, and they do not necessarily accurately show the actual spatial distribution of each vacant house.

8 Conclusion and Future Work

In this study, we developed a method to estimate the spatial distribution of vacant houses using public big data possessed by municipalities to implement and support vacant house distribution surveys in municipalities. We developed a vacant house

database to integrate public big data presumably possessed by many municipalities: BRR, HCI, and BRI, the digital RM, and field survey results in part of municipalities. Additionally, we were able to estimate the distribution of vacant houses with high reliability by utilizing the crosstab tables developed from the database. Our approach generated results that helped realize a method of conducting quick, inexpensive, and continuous surveys of vacant house distribution in a broad area, which had not yet been achieved by previous studies. Moreover, as mentioned in Sects. 1 and 2, social problems related to vacant houses occur not only in Japan but also in other countries worldwide. Since even municipalities outside Japan are expected to develop data similar to the public big data used in this study (of course, various other kinds of public big data are also usable), we believe our method can be applied outside Japan in the near future.

Based on this study, we also suggest the following directions for future research. First, it is necessary to apply our method to more municipalities and to study its possibilities and limitations. We only applied it to Kagoshima and Asakura; however, by applying it to various other municipalities, we could discover areas where its application is difficult, clarify its limitations, and find ways to improve the method. Second, it is expected that the estimation accuracy can be improved by proceeding with the collection of the field survey results on vacant house distribution and improving the information on the vacant house database. Third, this study used 15 kinds of variables, as shown in Table 5. However, we plan to increase this by collecting other kinds of municipal data since our method allows for greatly increasing the number of attributes. We hope to further increase the accuracy in consideration of the abovementioned scope for future work.

Acknowledgements This study received tremendous cooperation from the cities of Kagoshima and Asakura in Japan. We would like to thank them and all those people without whose cooperation this study would not have been realized. In addition, we would like to thank Editage (www.editage.jp) for English language editing.

References

- Abe S, Nakagawa D, Matsunaka R, Oba T (2014) Study on the factors to transform underused land focusing on the influence of railway stations in central areas of Japanese local cities. *Land Use Policy* 41:344–356
- Accordino J, Johnson GT (2002) Addressing the vacant and abandoned property problem. *J Urban Aff* 22(3):301–315
- Akiyama Y, Shibasaki R (2015) Future estimation of vacant house distribution using micro future population projection data. *CSIS DAYS 2015 research abstracts on spatial information science*, 41
- Asami Y (2014) Study of vacant land and vacant house in urban space. Progress, Tokyo
- Baba H, Asami Y (2017) Regional differences in the socio-economic and built-environment factors of vacant house ratio as a key indicator for spatial situation of shrinking cities. In: *Proceedings of 2017 international conference of Asian-Pacific planning societies*, #069
- Imai A, Sugimoto S, Sakakibara W, Mizuishi T (2015) Future of vacant house issue and possibility of utilization of used houses. *Knowl Creat Integr* 23(8):20–37

- Ishikawa D, Akagawa T (2014) Research on the evaluation method of vacant houses in Moji Ward Kitakyushu. *AJ Kyushu Chapter Arch Res Meet* 53:393–396
- Ishikawa M, Matsuhashi K, Kanamori Y, Ariga T (2017) Method of grasping detailed regional distribution of vacant dwelling based on the number of dwellings and households. *Trans CPIJ* 52(3):689–695
- Kato K, Yamamoto A, Kitajima T, Nakamura T, Nakashima H (2009) Emergence factors and renovation systems of vacant historical houses in a preservation district. *Mem Hous Res Found* 35:107–118
- Kubo T, Mashita M (2015) Geographical study of increase in housing vacancies in the central area of Gifu city. *Trans AJG* 247
- Nam J, Han J, Lee C (2016) Factors contributing to residential vacancy and some approaches to management in Gyeonggi province, Korea. *Sustainability* 8(4):367
- Sasaki T, Sano K, Kawabata M, Kaji M (2010) An intention and progress measures on the offer of vacant house in rural area. *J Rural Plan Assoc* 29(Special Issue):173–178
- Turnbull KG, Zahirovic-Herbert V (2010) Why do vacant houses sell for less: holding costs, bargaining power or stigma? *Real Estate Eco* 39(1):19–43
- Yamashita S, Morimoto A (2015) Study on occurrence pattern of the vacant houses in the local hub city. *Trans CPIJ* 50(3):932–937
- Zahirovich-Herbert V, Gibler KM (2014) The effect of new residential construction on housing prices. *J Hous Econ* 26:1–18

Enhancing the Use of Population Statistics Derived from Mobile Phone Users by Considering Building-Use Dependent Purpose of Stay



Toshihiro Osaragi and Ryo Kudo

Abstract Recently, it is possible to grasp the spatiotemporal distribution of people in cities using population statistics based on the location information of mobile phone users. However, it is difficult to know their purpose of stay which varies according to the use of building they stay and their detailed attributes such as age and gender. In this paper, we firstly propose a model that describes the number of people staying inside/outside of buildings by considering the population density that varies according to the use of building, time, and local characteristics, by using GIS database and Mobile Spatial Statistics (MSS) which is one of the population statistics of mobile phone users. Next, we integrate the MSS data and the Person Trip survey data (PT data) which include detailed personal attributes as well as the purpose of stay. Using the integrated database, we demonstrate the advanced use of population statistics based on mobile phone users by addition of purpose of stay which varies according to building use.

Keywords Mobile spatial statistics (MSS) · Person Trip survey data (PT data) · Multiple regression analysis · Spatiotemporal distribution · Population statistics

1 Introduction

Due to human activities and mobility by rapid urban transportation systems, the distribution of people varies according to time. This is especially true in metropolitan areas. The actual spatial distribution of transient occupants in any busy metropolitan area changes by the hour, or even by the minute, so static estimates of the population distribution are of limited utility (Osaragi 2009; Aubrecht et al. 2009, 2011). Concerning methods such as spatial disaggregation of population, one of the keys is to

T. Osaragi (✉) · R. Kudo

School of Environment and Society, Tokyo Institute of Technology, 2-12-1-M1-25 Ookayama, Meguro-ku, Tokyo 152-8550, Japan
e-mail: osaragi.t.aa@m.titech.ac.jp

R. Kudo

e-mail: kudo.r.ac@m.titech.ac.jp

© Springer Nature Switzerland AG 2020

P. Kyriakidis et al. (eds.), *Geospatial Technologies for Local and Regional Development*, Lecture Notes in Geoinformation and Cartography,
https://doi.org/10.1007/978-3-030-14745-7_11

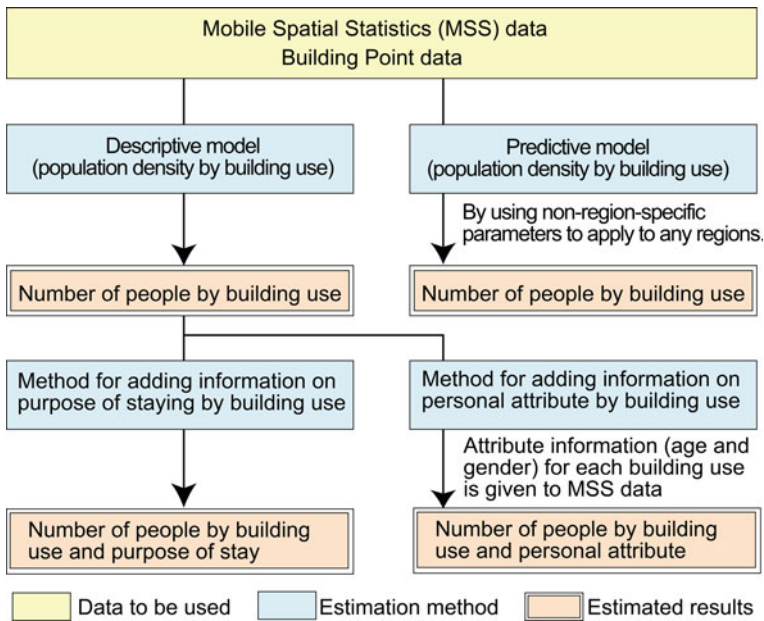


Fig. 1 Outline of the proposed model

combine such methods with earth observation data and remote sensing techniques in order to achieve fully integrated urban system models (Bracken and Martin 1989; Sim 2005; Steinnocher et al. 2006). With the increasingly available census data and remotely sensed data, to discuss their relationship is one of important issue in GIS data integration (Chen 1998, 2002).

Given this background, Osaragi and Hoshino (2012) proposed a statistical model to estimate the spatiotemporal distribution of people using Person Trip survey data (PT data), which are questionnaire-based survey conducted by the government, and include detailed personal attributes, as well as the location and time information of the departure and arrival, purpose of trip, and means of transportation (Osaragi 2015). However, PT data is compiled from the questionnaire survey conducted once every ten years and cannot provide the information on bank holidays. In addition, we cannot know the information of seasonal fluctuations and year-end and new-year holidays.

Recently, population statistics of mobile phone users that make it possible to determine the spatiotemporal distribution of people in urban areas are being put to practical use. However, detailed attribute information is sometimes concealed to protect privacy, and it is impossible to obtain specific information, such as the uses of the buildings where people are located or the purpose of stay in such buildings.

In this paper, we examine a method for enhancing the population statistics of mobile phone users by complementing the disadvantages and utilizing the advantages of urban demographic big data, by adding information on the purpose of stay which varies according to the uses of buildings where they stay.

First, we propose a model to estimate the number of people staying inside/outside of buildings for each building use, by using MSS data and the GIS data on the floor area and building use. Next, we integrate the data estimated by the above model and the Person Trip survey data (PT data). Furthermore, we propose a model that uses non-region-specific parameters so that it can be applied to anonymous regions by modifying the above model to describe the local characteristics of areas. Finally, using the estimated spatiotemporal distribution of people including detailed personal attributes and the purpose of stay, we attempt advanced use of the spatiotemporal distribution of people in urban areas. The outline of the proposed model is shown in Fig. 1.

2 Advanced Use of Population Statistics Based on Mobile Phone Users

2.1 Mobile Spatial Statistics (MSS) Data

Mobile Spatial Statistics (MSS) is one of the available population statistics based on mobile phone users, which was estimated by multiplying the number of mobile phones using the cellular phone network (NTT DoCoMo 2013) and the penetration rate (Oyabu et al. 2013) which is around 40% in Tokyo Metropolitan area (“Mobile Space Statistics” is a registered trademark of NTT DoCoMo, Inc.). The MSS data are provided as raster data of 500 m by 500 m grid cell (250 m by 250 m is available for densely populated area) for every hour. The largest advantage of the MSS data is that it is possible to grasp the precise spatiotemporal distribution of people every hour for all Japan area. As the disadvantage, the data do not provide any information to distinguish the people those who are staying inside building or those who are moving outside building. Some attributes such as age, gender, and location of residence are available, however, they are also sometimes concealed to protect privacy in case that the number of people is less than a certain number. Hence, we need to utilize the advantages of the MSS data and complement the disadvantages by using other available datasets.

2.2 Key Concept for Linking Population Statistics

Think about the people’s purposes of stays in each building and their attributes. In the case of amusement facilities, people’s purpose will be naturally to having time for amusement, and their attributes will be family with children or younger people. In

the case of fashionable commercial facilities, there will be relatively a larger number of young women visiting for shopping purposes. In the case of restaurants, people of diverse attributes visit there for eating and drinking. However, in the late night there are few families with children, and many of them will be office workers on weekdays. As shown in these examples, the purpose of stay and the attributes are deeply dependent on building use, day of the week, and time of day (Osaragi 2016).

Hence, first (1) we allocate the population data of MSS according to building use and time by using the model presented in the following section. Next (2) we calculate the composition ratios of the purpose and attributes of people for building use and time using PT data. Finally, (3) by multiplying the composition ratios to the number of people for each building use and time, and sum up them within each grid-cell, we can accurately estimate the people's purpose and attributes which greatly depend on building use and time. This is the key concept for linking MSS and PT data, which will be described in the following sections.

2.3 Estimation Method for Number of People by Each Building Use

First, we define a variable y_i^t , the total number of people staying inside/outside buildings in cell i at time t . Also, we define a variable y_{ij}^t , the number of people staying inside/outside buildings of building use j in cell i at time t . y_i^t is obtained by summing y_{ij}^t over all building uses j . Next, we consider that y_{ij}^t is proportional to the total floor area of building use j in cell i , denoted by x_{ij} [m^2]. The value of y_{ij}^t can be described by multiplying x_{ij} by the density of people, denoted by D_{ij}^t [persons/ m^2], which depends not only on building use j and time t , but is also dependent on the characteristics of cell i (location). Thus, y_{ij}^t can be described as follows;

$$y_i^t = \sum_{j=1} y_{ij}^t = \sum_{j=1} D_{ij}^t x_{ij}, \quad (1)$$

Next, we consider that the density of people, D_{ij}^t , can be decomposed into two parts; a time fluctuation factor α_j^t [person/ m^2] and a location characteristic factor β_i . A time fluctuation factor α_j^t is dependent on time t and building use j but common to all cells (location). A location characteristic factor β_i is dependent on each cell i but independent from time t and building use j . Thus, y_{ij}^t can be expressed as follows;

$$y_{ij}^t = \beta_i \sum_{j=1} \alpha_j^t x_{ij}. \quad (2)$$

When the data of population statistics on y_i^t and GIS data on x_{ij} are available, the unknown parameters α_j^t and β_i can be estimated by using multiple regression analysis. Herein, the model of the parameter describing local characteristics using the cell-specific unknown parameter β_i is referred to as the *descriptive model*.

On the other hand, in order to describe local characteristics, we propose the *predictive model* using the variables relating to the local characteristics as follows,

$$y_i^t = \left(\sum_{k=1} \gamma_k z_{ik} \right) \left(\sum_{j=1} \alpha_j^t x_{ij} \right), \tag{3}$$

where γ_k is the k th unknown parameter describing locality, and z_{ik} is the k th explanatory variable describing the local characteristics of the cell i .

The *descriptive model* uses a cell-specific parameter and might have a high estimation accuracy, but the geographical scope of application of the estimated model is limited. Meanwhile, when the *predictive model* uses general variables using available data, it can be applied to any regions for which the same kind of data can be obtained. Namely, the *predictive model* is useful for determining factors related to the number of people, while its accuracy might be lower than that of the *descriptive model*. Since we would like to validate the models and provide novel population dataset for Tokyo, where various spatial analyses have been done till date, we employ the descriptive model which is superior in terms of fitness.

2.4 Estimation Method for Number of People by Purpose of Stay

Person Trip survey has been carried out every ten years by the Ministry of Land, Infrastructure, Transport and Tourism of Japan, for detecting the actual travel behavior in cities. “Trips” defined in PT data are illustrated in Fig. 2. The information provided in PT data is shown in Table 1. Personal attributes (age, gender, occupation), the position and time information of departure and arrival, purpose of trip (18 purposes: e.g., commuting, business, shopping, eating), and means of transportation (5 means: on foot, bicycle, car, bus, train, ship, airplane) are included. The area for this survey covers a wide range of 70 km radius centered on the Tokyo Railway station. About 1.2 million persons were selected from around 33 million residents by random sampling based on census data. The number of valid samples is 883,044.

Using PT data, we can estimate the number of people with purpose l in building use j within cell i at time t , denoted by n_{ijl}^t . Although it seems be difficult to know the precise number of such people from PT data due to the limitation derived from time-interval of survey and low sample rate, we consider that the percentage of people with

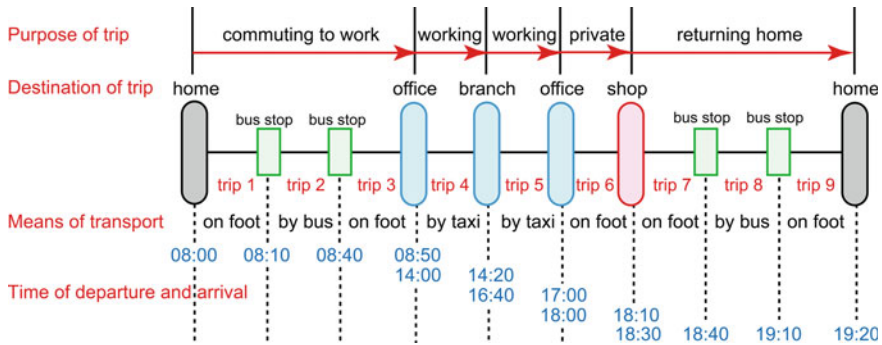


Fig. 2 Example of trips contained in Person Trip survey data

Table 1 Information provided in Person Trip survey data (PT data)

Item	Contents
Regions subject to survey	Tokyo, Kanagawa, Saitama, Chiba and Southern Ibaragi Prefectures
Survey time and day	24 h on weekdays in October of 1998, excluding Monday and Friday
Object of survey	Persons aged 5+ living in the above region
Sampling	Random sampling based on census data (1,235,883 persons selected from 32,896,705 persons)
Valid data	883,044 samples (mean weighting coefficient is approximately 37.3)
Content of data	Personal attributes, position and time of departure/arrival, purpose of trip, etc.
Purpose of trip	Purpose of each trip (18 purposes: e.g., commuting, business, shopping, eating)
Means of trip	Means of trip (5 means: on foot, bicycle, car, bus, train, ship, airplane)

purpose l in building use j within cell i at time t is relatively stable. This is because, people’s purpose of stay in the building of specific use in the specific location and time can be considered to be stable. Hence, the number of people with purpose l in building use j within cell i at time t , denoted by Y_{ijl}^t , can be expressed by using m_{ij}^t , the percentage of people with purpose l in building use j within cell i at time t , as follows,

$$Y_{ijl}^t = m_{ijl}^t y_{ij}^t, \quad \text{where } m_{ijl}^t = \frac{n_{ijl}^t}{\sum_{j=1} n_{ijl}^t}. \quad (4)$$

2.5 Estimation Method for Number of People by Building Use and Personal Attributes

The number of people of attribute u staying in cell i at time t , denoted by y_{iu}^t , is available from the MSS data. Using the PT data, it is possible to estimate the number of people of attribute u in building use j within cell i at time t , denoted by s_{iju}^t . As described above, we consider that the percentage of people of personal attributes u in building use j within cell i at time t , denoted by r_{iju}^t is stable. Thus, the number of people of attribute u in building use j within cell i at time t , Y_{iju}^t , is expressed by the following equation, so that the attribute composition percentage obtained from the MSS data is consistent.

$$Y_{iju}^t = \frac{r_{iju}^t y_{ij}^t}{\sum_{j=1} r_{iju}^t y_{ij}^t} y_{iu}^t, \quad \text{where } r_{iju}^t = \frac{s_{iju}^t}{\sum_{j=1} s_{iju}^t}. \tag{5}$$

3 Estimation Results for Number of People in an Urban Area

3.1 Details of Data Used in the Estimation

The explanatory variable, the total floor area x_{ij} , is constructed using Building Point data of FY2015 version (Zenrin Co Ltd.) possessing detailed building attribute information, which is aggregated according to building use j and cell i (cell size is 500 m by 500 m) (Table 2). The *predictive model* uses the number of buildings according to the size and use of buildings aggregated from the Building Point data (Table 3). In addition, the objective variable, the number of people y_i^t , is the mean values for the number of people in cell i at time t that is obtained from the MSS data. In order to analyze the differences of the day of the week, we analyze the data for weekdays and bank holidays respectively. The study area is shown in Fig. 3. The MSS data observed from 3:00 to 26:00 (107 weekdays and 47 bank holidays, 2015–2016) are used in the analysis. We used MSS whose cell size is 500 m by 500 m, because 250 m by 250 m is available only in high density areas (city centers).

3.2 Validation of Model

Figures 4 and 5 show the results for the number of people estimated by using the *descriptive model* and the *predictive model* respectively, according to weekday/bank holiday and time. The accuracy of the *descriptive model* is quite good, while underestimation or overestimation occurs in some of the cells in the *predictive model*. This

Table 2 Explanatory variable (total floor area by building use from building point data)

No.	Building use	Examples
1	Public/religious facility	Government office, Museum, Temple and shrine
2	Educational facility	Elementary/junior high/high school, University, Vocational school, study cram school
3	Medical/welfare facility	Hospital, nursing care welfare facility, etc.
4	Office	Logistics, Post office, Finance, Insurance, Real estate, Industry
5	Restaurant/bar	Home delivery, restaurant
6	Commercial facility	Service, sales business, automobile industry (including automobile repair business)
7	Mass retailer	Supermarket, department store, large store, etc.
8	Hotel	Hotel, Inn
9	Sports/entertainment facilities	Sports facility, entertainment
10	Apartment house	Flat, apartment house, dormitory, company house
11	Detached house	Detached house
12	Railway station	The number of railways (FY2005 National Land Numerical Data)
13	Park (outside buildings)	Area of park (Land Use Current State Survey Data, 2011)

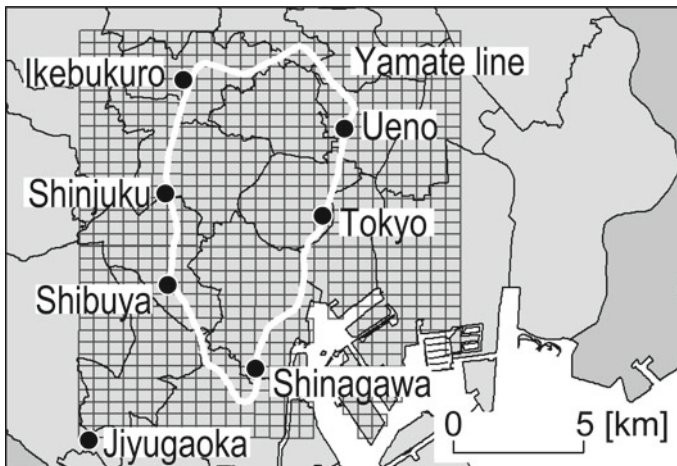


Fig. 3 Study area (500 m by 500 m grid cell, the total number of cells is 806)

Table 3 Number of buildings according to size (explanatory variables for the *predictive model*)

	No.	Building use	Size of building		
			Small	Medium	Large
Single building use	1	Public/religious facility	~150 m ²	150–600 m ²	600 m ² –
	2	Educational facility	~700 m ²	700–3300 m ²	3300 m ² –
	3	Medical/welfare facility	~175 m ²	175–350 m ²	350 m ² –
	4	Office	~80 m ²	80–210 m ²	210 m ² –
	5	Restaurant/bar	~75 m ²	75–150 m ²	150 m ² –
	6	Commercial facility	~105 m ²	105–225 m ²	225 m ² –
	7	Mass retailer	~190 m ²	190–360 m ²	360 m ² –
	8	Hotel	~550 m ²	550–1600 m ²	1600 m ² –
	9	Sports/entertainment facilities	~260 m ²	260–840 m ²	840 m ² –
	10	Apartment house	~200 m ²	200–500 m ²	500 m ² –
	11	Detached house	~80 m ²	80–160 m ²	160 m ² –
		12	Railway station	The number of railways (FY2005 National Land Numerical Data)	
	13	Park (outside buildings)	~330 m ²	330–1120 m ²	1120 m ² –
Compound building use	14	Commercial based complex (1)	~260 m ²	260–540 m ²	540 m ² –
	15	Commercial based complex (2)	~210 m ²	210–710 m ²	710 m ² –
	16	Office based complex (1)	~240 m ²	320–770 m ²	770 m ² –
	17	Office based complex (2)	~330 m ²	240–870 m ²	870 m ² –

result shows that the location characteristics are difficult to be explained by using the limited information on the size of buildings which are located in the cells. The other adequate explanatory variables should be incorporated in the model. This will be done in the future work.

Additionally, Fig. 6 shows the spatial distribution of the estimated parameter representing the value of location characteristic factor β_i for the *descriptive model* and the *predictive model*. Here, differences in location characteristics associated with the distribution of people can be confirmed. For example, a value of the *descriptive model* that is high in the central business district (CBD) around Tokyo Railway station on weekdays becomes low on bank holidays.

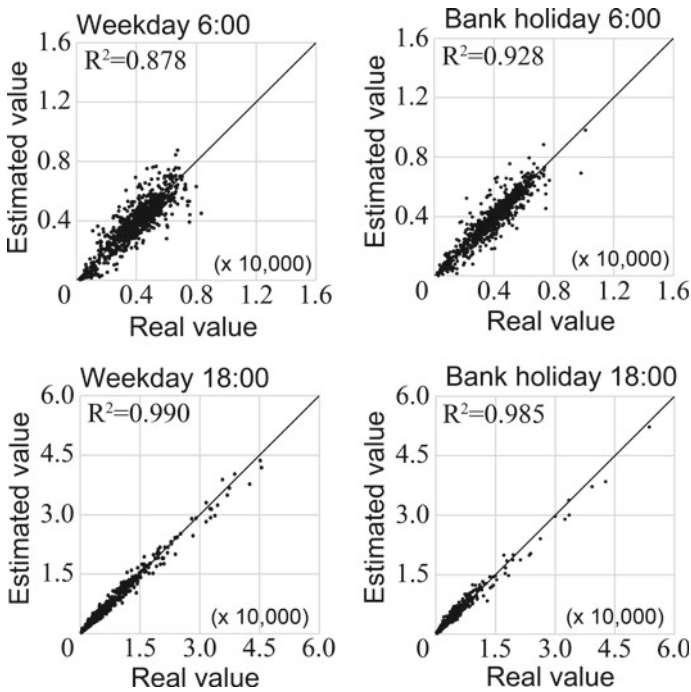


Fig. 4 Results for the number of people estimated by using the *descriptive model* according to weekday/bank holiday and time

Figure 7 shows the time fluctuation factor α_j^t , the temporal change in the density of the number of people by building use per 100 m², for the *descriptive model*. In apartment houses and detached houses, the rate of decrease in the density of people during the daytime is smaller on bank holidays than on weekdays. In addition, the time of peak density of people in restaurants/bars on weekdays differs from that seen on bank holidays. As a result, differences according to weekdays and bank holidays in building use can be confirmed.

3.3 Estimation Results for Number of People by Building Use and Purpose of Stay

Figure 8 shows the spatial distribution of the number of people according to building use at 12:00 on weekdays and on bank holidays. Here, it can be seen that people converge in (2) office buildings centered on Tokyo Railway station on weekdays, and in (4) commercial facilities centered on terminal stations such as Shibuya Railway station on bank holidays. Thus, we can break down the total number of people

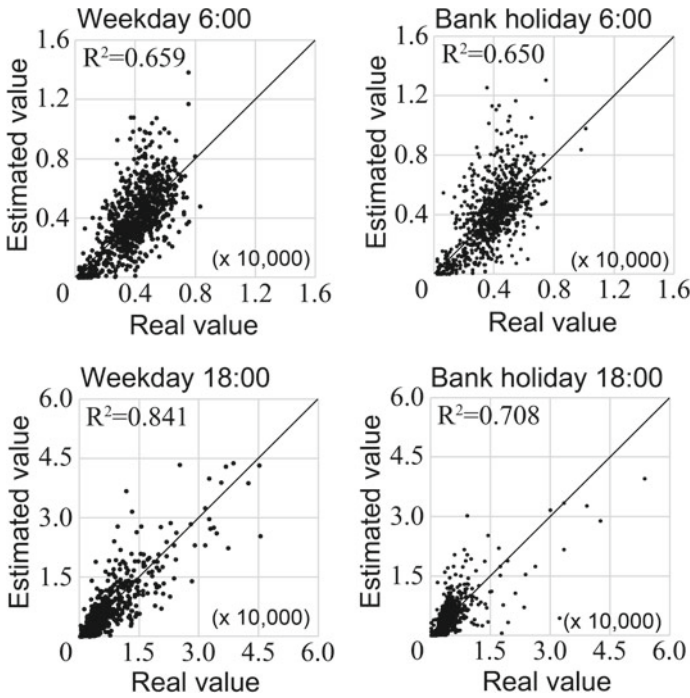


Fig. 5 Results for the number of people estimated by using the *predictive model* according to weekday/bank holiday and time

obtained from the MSS data into the number of people who are staying inside/outside building for each specific building use. This is a novel aspect of the proposed method in this research.

Furthermore, Fig. 9 shows the temporal change in the number of people by building use and the purpose of stay in the cell where JR Ueno Railway station is located. Looking at the number of people inside/outside the station building on a bank holidays by purpose (Fig. 9b right), it can be seen that there are many people with leisure as a purpose at around 11:00, while there are many people who are shopping or returning home between 16:00 and 19:00. Ueno railway station is one of the hub railway station in Tokyo, and it is well known that people with various kinds of purposes are passing through. We can see that dynamic change of people’s purpose of trips which varies according to the day of week.

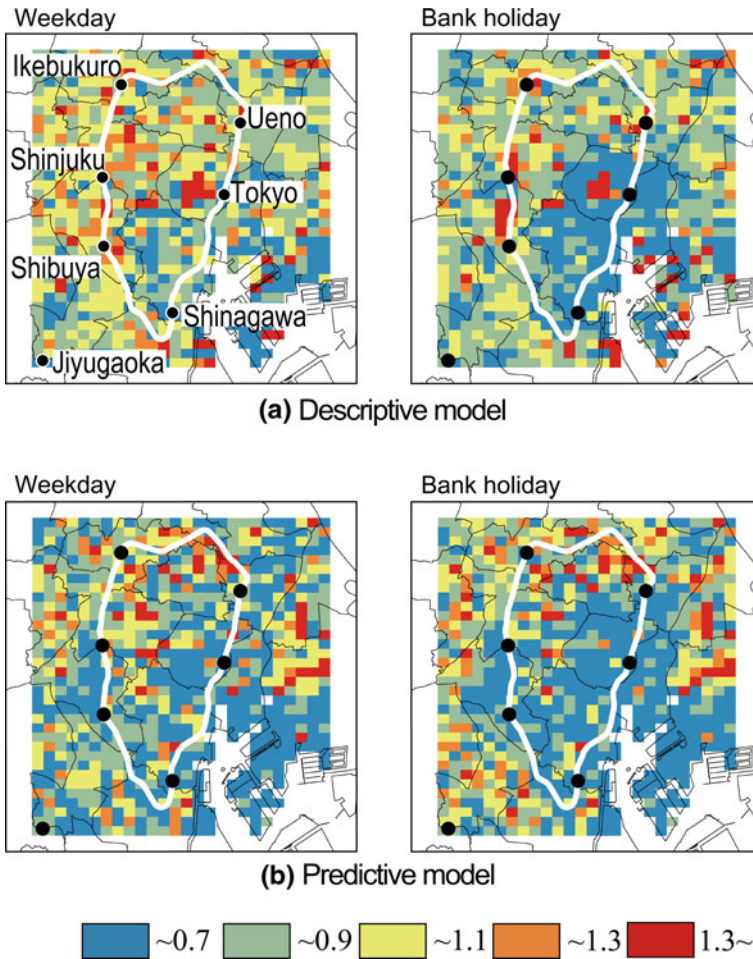


Fig. 6 Spatial distribution of the estimated parameters representing location characteristic factor β_i for the descriptive model and the predictive model

3.4 Estimation Results for Number of People by Building Use and Personal Attributes

Figure 10 shows the spatial distribution of the number of people in commercial facilities at 18:00 by attributes. Here, we can confirm that the cells where people are concentrated differ according to people’s attributes. For example, many men and women in their twenties and thirties in the vicinity of Shibuya railway station, where is very popular for younger generations. Also, men and women in their thirties to fifties are concentrated in the cells around Asakusa railway station, where is one of the typical down town areas in Tokyo.

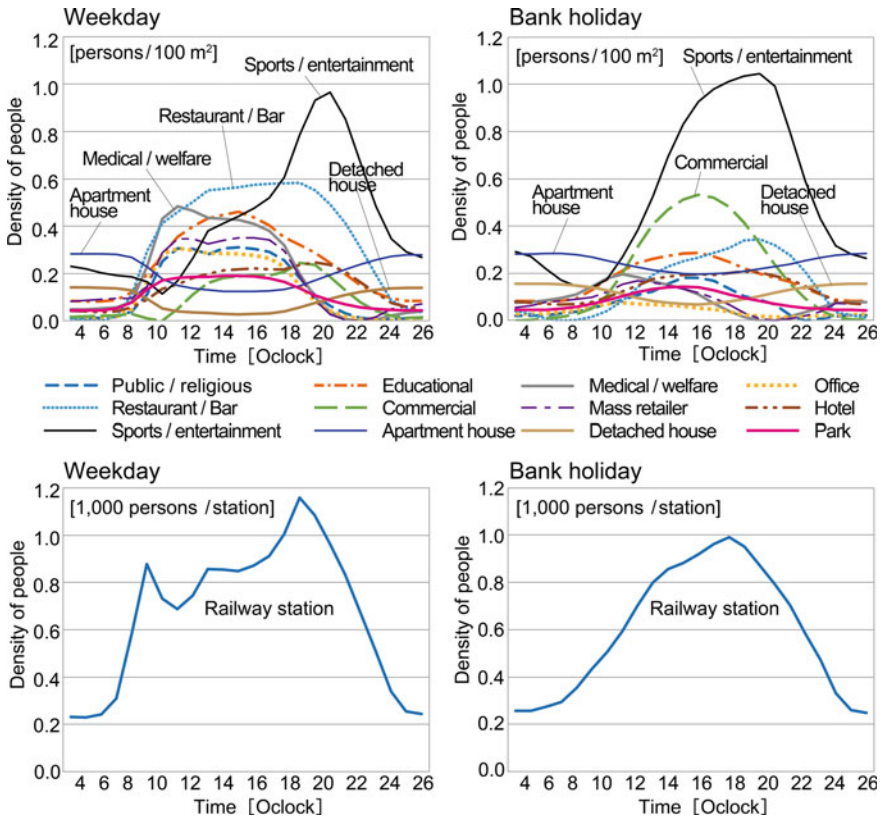
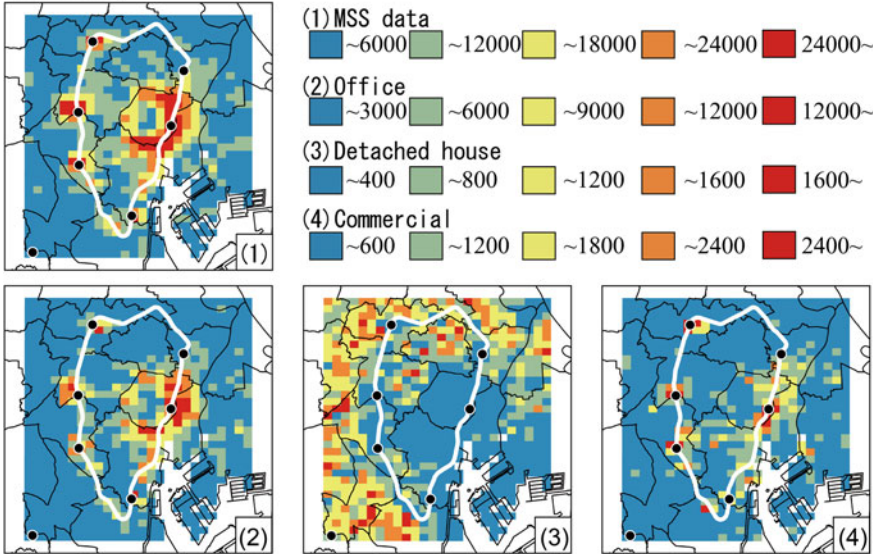


Fig. 7 Temporal change in the density of number of people by building use per 100 m², for the descriptive model

Figure 11 shows the temporal change in the number of people by building use and personal attributes in the specific cells. Here, the differences in the attributes of people according to location characteristics are apparent. For example, during the day there are many men and women in their twenties and thirties in the cell containing the Ginza railway station which is the commercial center of Tokyo and very popular for younger generations. Also, there are many men in their twenties and thirties in the cell containing Akihabara railway station where many electric stores are located and famous for that many foreigners visit.

These numerical examples demonstrate that advanced use can be possible for the MSS data which provide only the total number of people.

Weekday



Bank holiday

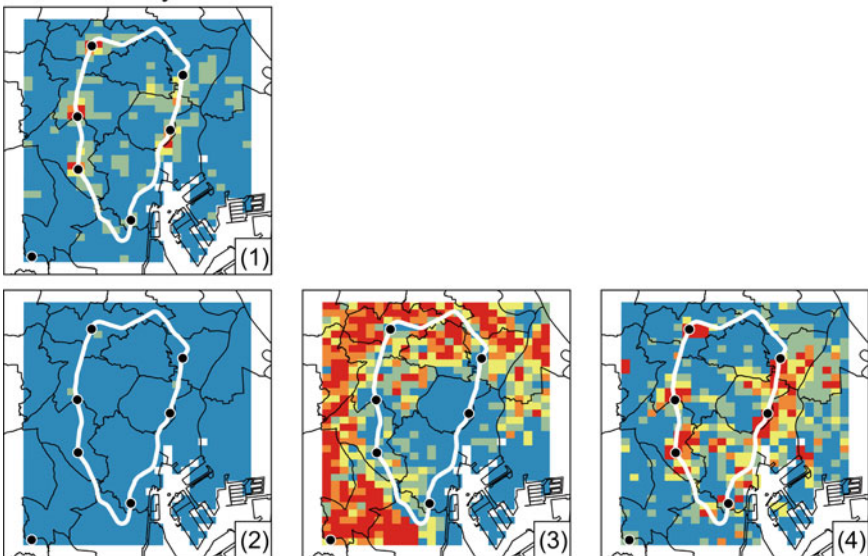


Fig. 8 Spatial distribution of the number of people according to building use at 12:00 on weekdays and on bank holidays

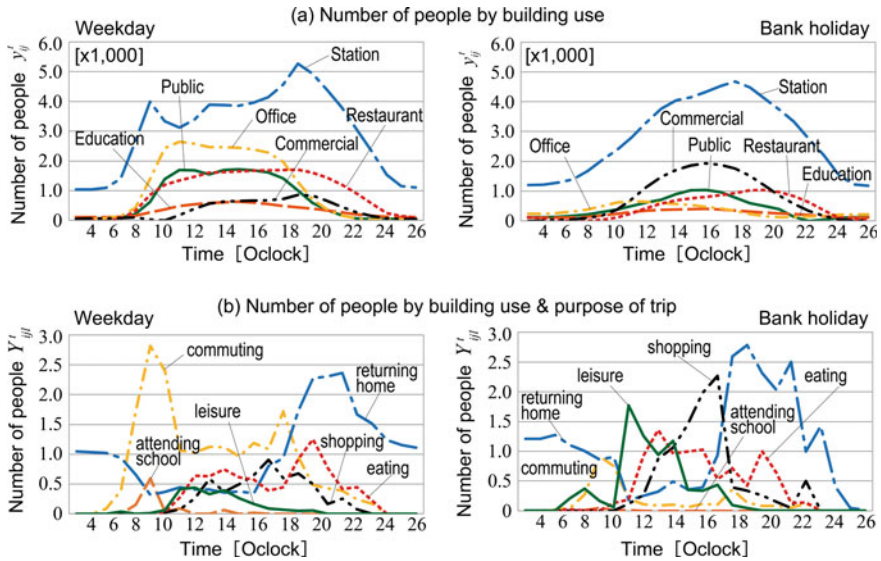


Fig. 9 Temporal change in the number of people by **a** building use **b** purpose of stay in the cell where JR Ueno Station is located

4 Summary and Conclusions

We introduced the population statistics based on the location information of mobile phone users, and pointed out its limitation in which detailed personal attribute information is being concealed to protect privacy. Given this back ground, we proposed a method for enriching the value of the population statistics based on mobile phone users. More specifically, we added the detailed information on the attributes (age, gender) and the purposes of their stays, by focusing on the uses of buildings where they are staying, as follows.

(1) First, we proposed a model to estimate the number of people being inside/outside of buildings for each building use in detailed units of space and time, by using Mobile Spatial Statistics (MSS) data and Building Point data, which is the GIS data including detailed attribute information such as floor area and building use. (2) Next, we added the detailed attribute information to the people being inside/outside of buildings for each building use, by using the Person Trip survey data (PT data) in which detailed personal attributes as well as the location and time information of the departure and arrival, purpose of trip, and means of transportation are included. In this process, we utilized the fact that detailed personal attributes of people are deeply dependent on the time, location, and building use where they are staying. (3) Finally, using the spatiotemporal distribution of people which was estimated by the proposed model, we attempted to making regional comparisons of the spatiotemporal distribution of people in urban areas. Namely, we demonstrated that it is possible to grasp the spatiotemporal characteristics of population distribution

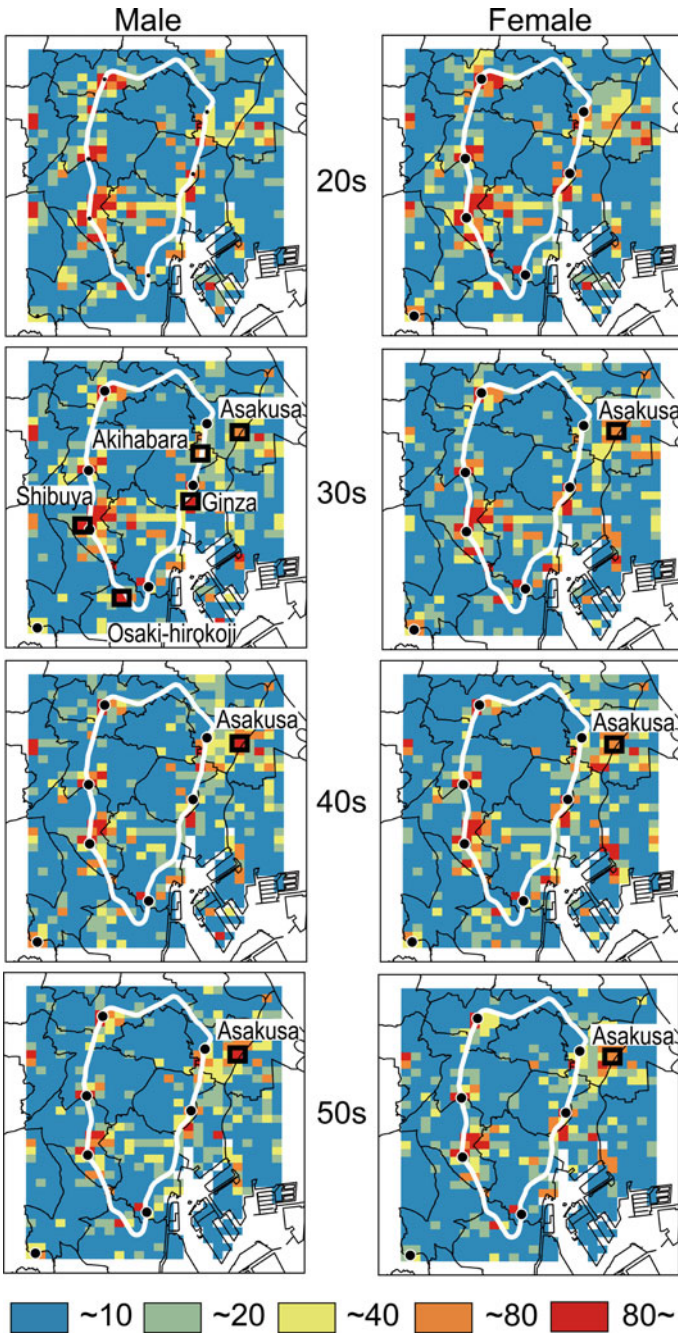


Fig. 10 Spatial distribution of the number of people in commercial facilities at 18:00 by attributes

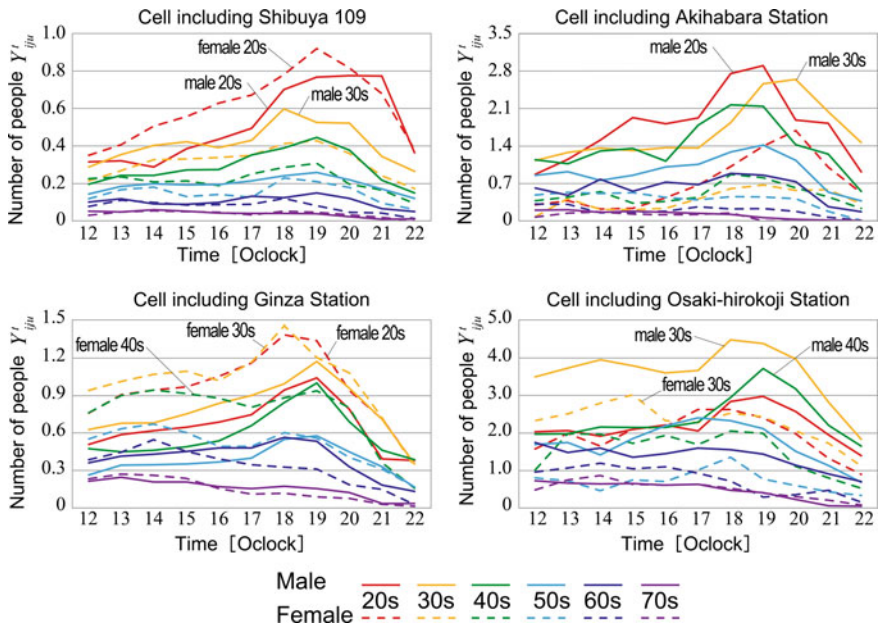


Fig. 11 Temporal change in the number of people by building use and personal attributes in specific cells

attached with detailed attributes information such as their age, gender, and purposes that vary according to the time, location, and building use where they are.

There is a growing demand for data that allows highly accurate understanding of the spatiotemporal distribution of both moving and static people in urban areas. The method proposed in this paper, however, cannot provide such information. In order to address this issue, we would like to construct a method of estimating the number of people who flow in or flow out in detailed units of space and time, by integrating multiple sets of data. Furthermore, we would like to discuss the characteristics of spatiotemporal distribution of moving people that varies according to regional characteristic, day of week, and time.

Using the models proposed in this paper, we can enrich the value of existing population statistics. The models are, however, basically assuming the use of the specific data available in Japan. In order to expand the availability of the models for other data sources available in other countries, we need to discuss differences and commonalities in population statistics, which are currently available in other countries (Ahas et al. 2015; Gao 2015; Järv et al. 2017).

Acknowledgements This paper is part of the research outcomes funded by KAKENHI (Grant Number 17H00843). The authors wish to express their sincere thanks for valuable comments from anonymous reviewers of AGILE 2019.

References

- Ahas R, Aasa A, Yuan Y, Raubal M, Smoreda Z, Liu Y, Ziemlicki C, Tiru M, Zook M (2015) Everyday space–time geographies: using mobile phone-based sensor data to monitor urban activity in Harbin, Paris, and Tallinn, Issue 11: Geographies of Mobility: Applications of Location Based Data. *Int J Geogr Inf Sci* 29:2017–2039
- Aubrecht C, Steinnocher K, Hollaus M, Wagner W (2009) Integrating earth observation and GIScience for high resolution spatial and functional modeling of urban land use. *Comput Environ Urban Syst* 33:15–25
- Aubrecht C, Köstl M, Steinnocher K (2011) Population exposure and impact assessment: benefits of modeling urban land use in very high spatial and thematic detail. In: *Computational vision and medical image processing computational methods in applied sciences*, vol 19. Springer, Dordrecht, pp 75–89
- Bracken I, Martin D (1989) The generation of spatial population distributions from census centroid data. *Environ Plann A* 21:537–543
- Chen K (1998) Correlations between census dwelling data and remotely sensed data. In: *Proceedings: SIRC 98—10th annual colloquium of the spatial information research centre*, Dunedin, New Zealand
- Chen K (2002) An approach to linking remotely sensed data and areal census data. *Int J Remote Sens* 23(1):37–48
- Gao S (2015) Spatio-temporal analytics for exploring human mobility patterns and urban dynamics in the mobile age. *Spat Cogn Comput* 15(2):86–114
- Järvi O, Tenkanen H, Toivonen T (2017) Enhancing spatial accuracy of mobile phone data using multi-temporal dasymetric interpolation. *Int J Geogr Inf Sci* 31:1630–1651
- NTT DoCoMo (2013) NTT DoCoMo technical journal. https://www.nttdocomo.co.jp/english/corporate/technology/rd/technical_journal/bn/vol14_3/. Accessed 26 Jan 2019
- Osaragi T (2009) Estimating spatio-temporal distribution of railroad users and its application to disaster prevention planning. In: Sester M et al (eds) *Lecture notes in geoinformation and cartography, advances in GIScience*. Springer, Berlin, pp 233–250
- Osaragi T (2015) Spatiotemporal distribution of automobile users: estimation method and applications to disaster mitigation planning. In: *12th international conference on information systems for crisis response and management (ISCRAM 2015)*, proceedings of the ISCRAM 2015 conference, ISCRAM 2015 organization, pp 87–99
- Osaragi T (2016) Estimation of transient occupants on weekdays and weekends for risk exposure analysis. In: *13th international conference on information systems for crisis response and management (ISCRAM 2016)*, proceedings of the ISCRAM 2016 conference, ISCRAM 2016 organization
- Osaragi T, Hoshino T (2012) Predicting spatiotemporal distribution of transient occupants in urban areas. In: *15th AGILE conference on geographic information science, lecture notes in geoinformation and cartography, bridging the geographic information sciences*. Springer, Berlin, pp 307–325
- Oyabu Y, Terada M, Yamaguchi T, Iwasawa S, Haiwara J, Koizumi D (2013) Evaluating reliability of mobile spatial statistics. *NTT DoCoMo Tech J*. https://www.nttdocomo.co.jp/english/binary/pdf/corporate/technology/rd/technical_journal/bn/vol14_3/vol14_3_016en.pdf. Accessed 26 Jan 2019

- Sim S (2005) A proposed method for disaggregating census data using object-oriented image classification and GIS. *Int Arch Photogramm Remote Sens Spat Inf Sci XXXVI(Part 8/W27)*
- Steinnocher K, Weichselbaum J, Köstl M (2006) Linking remote sensing and demographic analysis in urbanised areas. In: Hostert P, Damm A, Schiefer S (eds) *Proceedings: first workshop of the EARSeL SIG on urban remote sensing*, Berlin, Germany

Potential of Crowdsourced Traces for Detecting Updates in Authoritative Geographic Data



Stefan S. Ivanovic, Ana-Maria Olteanu-Raimond, Sébastien Mustière and Thomas Devogele

Abstract Crowdsourced traces collected by GPS devices during sports activities are now widely available on different websites. The goal of this paper is to study the potential of crowdsourced traces coming from GPS devices to highlight updates in authoritative geographic data. To reach this goal, an approach based on two steps is proposed. First, a data matching method is applied to match authoritative data and crowdsourced traces. Second, for the non-matched crowdsourced segments composing a trace, different criteria are defined to decide if whether or not, non-matched segments should be considered as an alert for update in authoritative data. The proposed approach is tested on crowdsourced traces and on BDTOPO® authoritative road and path network in mountain area. The results are promising: 727, 1 km of missing paths were found in the test area, which corresponds to 7.7% of the total length of used traces. The discovered missing paths also represent a contribution of 2.4% of the total length of BDTopo® road and path network in the test area.

Keywords Data matching · Crowdsourced GPS traces · Detection of updates · Authoritative geographic data · Decision making

S. S. Ivanovic · A.-M. Olteanu-Raimond (✉) · S. Mustière
University of Paris-Est, LASTIG MEIG, IGN, ENSG, F-94160 Saint-Mandé, France
e-mail: ana-maria.raimond@ign.fr

S. S. Ivanovic
e-mail: stefangrf@gmail.com

S. Mustière
e-mail: sebastien.mustiere@ign.fr

T. Devogele
Université François Rabelais de Tours, Tours, France
e-mail: thomas.devogele@univ-tours.fr

© Springer Nature Switzerland AG 2020
P. Kyriakidis et al. (eds.), *Geospatial Technologies for Local and Regional Development*, Lecture Notes in Geoinformation and Cartography,
https://doi.org/10.1007/978-3-030-14745-7_12

1 Introduction

Nowadays, in the era of open data and crowdsourced data, the need for very up to date authoritative geographic data has significantly increased. Generally, authoritative data are updated continuously or regularly, mainly through stereo-restitution (i.e. 3D mapping from aerial and satellite images) and field surveys. This is a very time-consuming and expensive task (Olteanu-Raimond et al. 2016). The road and path network is an especially challenging theme regarding updates in reason of the frequent changes of these objects, especially for footpath, tractor or bicycle paths due to their intermittent nature (e.g. they appear and disappear very often) and their location in various landscapes sometime difficult to survey (e.g. forest, high mountains, seashore). Using stereo-restitution techniques is challenging in mountainous area where dense vegetation may limit visibility from the sky. Field surveys are very expensive due to the vast open area that has to be covered. Thus, these types of roads, named paths in this paper, are less up-to-date in authoritative data compared to main roads such as highway and national roads.

Simultaneously, thanks to developments of Web 2.0 and the integration of GPS (Global Positioning Systems) or other GNSS (Global Navigation Satellite Systems) into smartphones or other portable devices, citizens may now act as sensors and generate geographic data, becoming ‘producers’ as tackled by Bruns (2009). The wider use of terms such as crowdsourced data, volunteer geographic information (VGI), citizen-science, and user generated content (UGC) refer to the increasing involvement of citizens in geographic data collection (See et al. 2016). Crowdsourced traces are thus collected by GPS devices during sport and spare time activities, with various devices from very low to high class. The question of their usability is then open. On the one hand, traces are collected without any protocol, with low and heterogeneous frequency sampling, and are made available with few or inexistent metadata. Moreover, various errors are introduced by different uncontrolled external factors such as the position of the device (in bags or pockets...) or the canopy cover. These errors can cause significant errors and may limit the usability of the traces. On the other hand, the precision of these traces is continuously increasing and their interest is shown in the literature for purposes such as bicycle routing (Bergman and Oksanen 2016) or updating or enriching main roads from authoritative data (Al-Bakri and Fairbairn 2012; Liu et al. 2015; van Winden et al. 2016).

The goal of this paper is to study the potential of crowdsourced traces for updating authoritative data, and especially for unpaved paths. Despite the limitation of the crowdsourced traces as discussed below, our assumption is that these traces can be used to highlight missing paths in authoritative data, or confirm that some paths are still in use. To reach the goal, a two-steps approach is proposed. The first step consists in matching traces with the authoritative road network to identify segments of traces which have no homologous counterpart in the authoritative data. The second step, named decision making, consist in defining if a non-matched segment should be considered as an alert for update.

The paper is organized as follows. First, Sect. 2 describes the state of the art concerning the use of crowdsourced data for updating authoritative data or more globally in data matching processes. Section 3 describes the proposed data matching approach for identifying potential missing paths in authoritative data. The proposed method for highlighting actual updates in authoritative data is described in Sect. 4. Before concluding, the results are presented in Sect. 5.

2 State of the Art

The use of crowdsourced data for update or enrichment purposes has been considered in many research works. Al-Bakri and Fairbairn (2012) concludes that despite the richness of crowdsourced data, there are significant incompatibilities, especially the accuracy, in integration of OpenStreetMap (OSM) data into authoritative data from Ordnance Survey. On the other hand, Zielstra and Hochmair (2011) found satisfying the integration of OSM data with data coming from TeleAtlas and NAVTEQ. However, this refers to a specific context where OSM pedestrian paths are considered useful for enhancing accessibility to bus and metro stations in US and German cities. van Winden et al. (2016) used crowdsourced traces collected for research purposes to update Dutch authoritative road characteristics (e.g. one-way or two-way road). Liu et al. (2015) proposed an approach to detect new roads for updating authoritative road network by using OSM data, which requires overcoming incompatibilities between data. Other research works claim that the positional accuracy and completeness (Girres and Touya 2010; Koukoletsos et al. 2012), actuality (Jokar Arsanjani et al. 2013) and semantic accuracy (Fan et al. 2014) of much crowdsourced data may actually be sufficient for the needs of mapping agencies, even if there exist issues regarding the heterogeneity and inconsistency of crowdsourced data. As mentioned in Olteanu-Raimond et al. (2016), National Mapping Agencies (NMA) have started to engage with crowdsourced data. To maintain their reputation for high quality products, they proposed their own crowd and community platforms to update their data (Olteanu-Raimond et al. 2016, 2018). Let us precise that, in this work we are only interested only in detecting updates for authoritative data, the process of integrating those updates being left to the NMA.

In the cited works, data matching is a tool for defining homologous features between crowdsourced and authoritative data. Many data matching methods rely on the comparison of geometric positions, thematic and semantic information, and an analysis of relations between features, or a combination of all those aspects. A detailed state of the art on data matching is described in Olteanu-Raimond et al. (2015), Xavier et al. (2016). When dealing with traces from GPS devices, and especially for navigation purposes, GPS location points are assigned to most probable road within a road network. This process is a special case of data matching known as map-matching, which received a much attention in the literature (Lou et al. 2009; Newson and Krumm 2009).

Despite their advantages most current matching methods are not adapted in our case. First, data matching methods relying on topology (Mustière and Devogele 2008) are not appropriate in our case, since those methods usually expect a relatively complete network, which is not the case of network built from traces. Second, methods based on thematic and semantic comparisons (e.g. Olteanu-Raimond et al. 2015), are not eligible due to lack of metadata and thematic and semantic information in GPS traces. Finally, besides our data are coming from GPS devices, the map matching solutions (e.g. Newson and Krumm 2009) are not appropriate due to the purpose of our map matching method focus on matching complete GPS traces to existing roads, while we are especially interested in detecting cases of not-matched parts of traces.

We thus expect that a fully-geometric based method using buffer growing (Walter and Fritsch 1999; Liu et al. 2015) is more adapted in our case. However, existing buffer based approaches are using single fixed thresholds (Liu et al. 2015). In case of heterogeneous traces, applying a single threshold can cause significant side effects due to the heterogeneity of precision of traces (i.e. false positive and negative matchings).

3 Data Matching for Identifying Missing Paths

This section describes our data matching approach to identify potential missing paths in authoritative data.

3.1 Data Matching Approach

In this research, data matching is used as a tool for detecting potential updates in authoritative data by identifying differences between two datasets: crowdsources traces (noted DS_1) and authoritative data (noted DS_2). The results of a data matching are a set of links defining homologous features and having different cardinalities: 1:0, n:m, and 0:1. In this work, we are interested in matching links having the cardinality 1:0 (i.e. one feature from DS_1 does not have a corresponding feature in DS_2). Our assumption is that features in crowdsourced traces that are not matched to features from authoritative dataset are considered as candidates for potential updates.

The proposed data matching approach consists of two steps: selection of candidates to match and analysis of candidates.

3.1.1 Matching: Selection of Candidates

The first step is the selection of candidates. Each trace from DS_1 is composed of a set of segments $\{S_{\text{trace},i}\}$, where $i = 1, \dots, M$; M represents the total number of segments composing the trace. It is assumed in this approach that GPS traces are

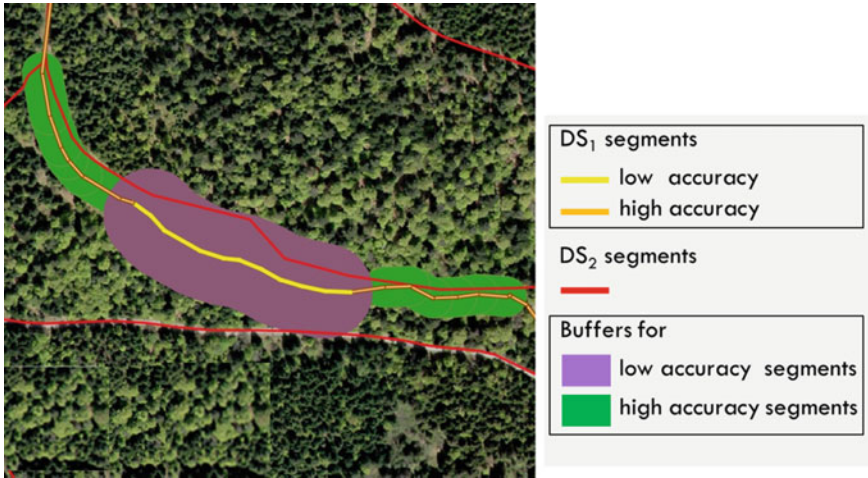


Fig. 1 Buffer for selecting candidates to match based on the accuracy of traces’ segments

associated with an a priori evaluation of their accuracy, at least a binary evaluation as probably ‘low’ or ‘high’ accuracy. This may originate from metadata of traces in the best possible circumstances. This may also originate from a pre-evaluation of this accuracy based on an intrinsic and extrinsic analysis of the trace as proposed by the authors: a chaotic shape of the trace or a complex environment like forest or town area may be indicator of probable low accuracy of traces (Ivanovic et al. 2015).

For each segment $S_{trace,i}$, composing a trace from DS_1 , we search for close segments in DS_2 , according to a distance criterion implemented by buffers. Let us consider $\{S_{Auth,j}\}, j = 1, \dots, N$, the set of segments from DS_2 that intersect the buffer of the segment $S_{trace,i}$. These selected segments are the candidates for matching with $S_{trace,i}$. Knowing the heterogeneity of the accuracy inside the trace, an adaptable buffer depending on the segment accuracy is proposed. Hence, a greater size is used for segments having ‘low’ accuracy and a smaller size for segments having ‘high’ accuracy. The selection strategy is illustrated in Fig. 1 where purple and green buffers are generated respectively for low and high accuracy segments.

3.1.2 Matching: Analysis of Candidates

The second step of the matching approach is to analyse each candidate to define the actual matching links between homologous features. Thus, three criteria are defined to compare the segment to match ($S_{trace,i}$) and the candidates: angle, neighbourhood, and length. The criteria are applied one after the other: the angle criterion is first applied followed by the neighbourhood criterion. Finally, after merging the contiguous matched and respectively unmatched segments obtained, the length criterion is applied.

Direction criterion. This criterion is based on the assumption that collinear features are more likely homologous features than perpendicular features (Olteanu-Raimond et al. 2015). The criterion compares local orientations between $S_{\text{trace},i}$ and a candidate for matching, $S_{\text{Auth},j}$. The orientation is evaluated through the angular difference q between the orientations of tangents to $S_{\text{trace},i}$ and to $S_{\text{Auth},j}$ respectively where the point of $S_{\text{trace},i}$ is nearest to $S_{\text{Auth},j}$, and where the point of $S_{\text{Auth},j}$ is nearest to $S_{\text{trace},i}$. If the angle between the two segments is less than a threshold, T_D , segments are considering as corresponding segments.

Neighborhood criterion. This criterion is based on the assumption that the matching of on feature depends on the matching of its neighbours (Olteanu-Raimond et al. 2015). Hence, for each segment, $S_{\text{trace},i}$, the algorithm analyses if the preceding and succeeding segments have the same matching result (matched or not matched). More precisely, the two neighbouring preceding and the two succeeding segments are taken into account. If all neighbour segments have different matching results than the current candidate, the result is changed for the candidate.

Length criterion. The assumption for this criterion is that for a valuable update, the segment should have a minimal length, in order to respect data specifications, and in order to avoid insignificant update considered as noise. The criterion is defined as follows. First, initial $S_{\text{trace},i}$, segments are merged according to their matching results (matched or unmatched) obtained after applying Direction and Neighborhood criteria. For each unmatched aggregated segment, the algorithm verifies if the length of the segment is higher than a threshold, T_L . If it is the case, then the segment is classified as unmatched.

4 Decision Making

Determining which unmatched segment represents a real update is a difficult task. Different interpretations are possible for unmatched segments, as illustrated for example in the special case of a “deviation” in Fig. 2: (i) GPS error: the deviation of a part of the trace from authoritative road is caused by GPS error measurements; (ii) New road: the deviation is following an existing road in the real world which is not represented in the database; (iii) Road modification: the deviation is due to a modification of the existing road. The last two cases are considered as missing roads, thus potential and different updates.

We propose to combine different criteria to define a degree of confidence measuring to which extent the unmatched segment may be proposed for an update.

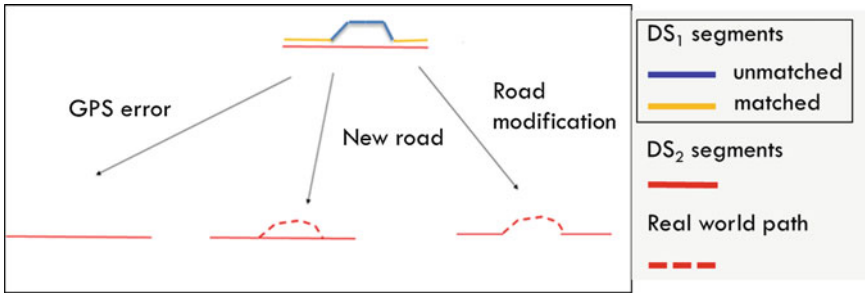


Fig. 2 Different interpretations of the unmatched segments of a trace

4.1 Definition of Criteria

Taking into account the nature and specificity of traces, four relevant criteria are proposed: quantity, accuracy, actuality, and continuity. For each criterion a score of confidence is computed.

4.1.1 Quantity Criterion

The criterion refers to the number of traces representing the same missing path and is based on the following idea: the more non-matched segments belonging to different traces follow the same path exist, the more chances that this path exists in the reality and is missing in the database.

To determine if segments of different traces are representing the same missing path, a buffer is first computed for each unmatched segment. All other unmatched segments having more than a certain percentage of their lengths within this buffer are considered as following the same missing path. Second, directions between the candidates are computed as proposed by Olteanu-Raimond et al. (2015). Only the candidates having similar directions are finally considered as following the same missing path. The scores for the quantity criterion C_1 are computed as follows:

1. Rule 1: if only one candidate for a missing path exists, then criterion score is equals to 0.25.
2. Rule 2: If two candidates following the same path exist, then criterion score is 0.5.
3. Rule 3: If more than two candidates follow the same path, then criterion score is 1.

Due to the weak redundancy of traces in mountainous areas, only three rules are defined. Nevertheless, the computation of the scores can be modified in case of higher redundancy of traces, by introducing more rules and modifying threshold regarding the expected redundancy of traces.

4.1.2 Accuracy Criterion

This criterion is based on the accuracy assessment of segments composing a trace. Our hypothesis is that an unmatched segment with high accuracy is more reliable than one segment with low accuracy.

Even if data matching already takes into account the accuracy of segments of traces, there are still situations where low accuracy segments can negatively bias the results of data matching.

The accuracy criterion score, C_2 , is computed for each unmatched segment as the percentage of length of high quality segments in its total length.

4.1.3 Actuality Criterion

For the detection of updates, the actuality of traces is considered as an important factor. The assumption for this criterion is that the more recent a trace is, the more relevant the update will be. In our context, we only consider the year in which a trace is imported on the website, which is sometimes the only available information. The confidence scores for the actuality criterion, C_3 , are defined as follows:

1. Rule 1: If the trace is collected during the current year, the confidence score equals 1.
2. Rule 2: If the trace is collected during four previous years the score is reduced by 0.25, being equal to respectively: 0.75, 0.5 and 0.25.
3. Rule 3: If the trace is collected four years ago and before, then the confidence score equals zero.

4.1.4 Continuity Criterion

Cases where a trace follows an existing authoritative path along most of its length except in some parts may occur. Some of these cases should not be considered as updates, and are due to the heterogeneities regarding the accuracy of segments for the same trace or alternatively to human behaviour living regular roads (Fig. 3, on the left). Sometimes these cases should be considered as updates, due to the fact that the trace really follows a path that is not represented in authoritative data (Fig. 3, on the right). In the following, the first case is named ‘unmatched segment following the same authoritative path’ and the second case is named ‘unmatched segment not following the same authoritative path’.

To distinguish these two situations, the continuity criterion is defined as follows. First, unmatched segments whose neighbours are following the same authoritative path are identified, $\{S_{\text{unmatched},i}; i = 1, \dots, M\}$, where M is the number of selected unmatched segments. Second, for each unmatched segment, $S_{\text{unmatched},i}$, its length and the distance to the nearest authoritative path are computed to distinguish between real missing paths and deviations due to low accuracy and behaviour. The nearest

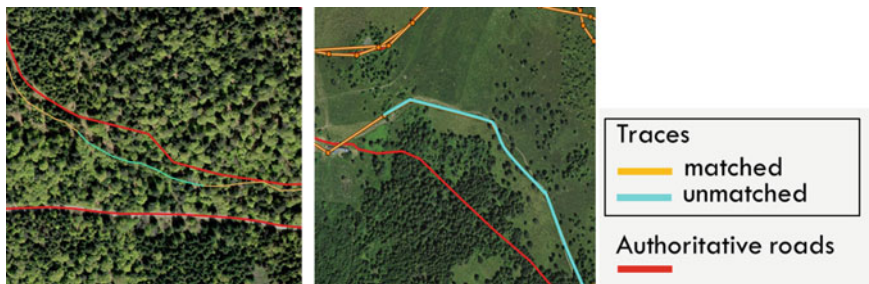


Fig. 3 Part of the trace not following authoritative roads: due to human behaviour or GPS error (on the left); due to a missing path in authoritative data (on the right)

authoritative path is selected by defining a growing buffer for each $S_{unmatched,i}$, i.e. buffer is enlarged until the first authoritative path is selected. The score confidence of continuity criterion, C_4 , is defined as follows:

1. Rule 1: If $(length(S_{unmatched,i}) < T_L \text{ and } d_H < T_d)$, then the score equals 1, otherwise the score confidence equals 0.

where: $length(S_{unmatched,i})$ is the length of the segment, $S_{unmatched,i}$; T_L the threshold determining ‘short parts’; d_H is Hausdorff semi-distance between unmatched segment and the nearest authoritative path, and T_d the threshold determining the closeness between an unmatched segment and the nearest authoritative path.

4.2 Combination of Criteria

The four criteria are combined to compute a degree of confidence for each unmatched segment, as shown in Eq. 1:

$$CD_i = \sum_{j=1}^n C_{ij} W_j \tag{1}$$

where: CD_i is the confidence degree for i -th unmatched segment, and C_{ij} is the normalized score of i -th unmatched segment with the respect to j -th criterion and W_j is the weight of criteria j .

Based on the confidence degree, each unmatched segment is classified into three classes:

- Low confidence if $S_i \leq T_1$. We weakly believe that an unmatched segment is a missing path.
- Medium confidence if $T_1 < S_i \leq T_2$. We moderately believe that the unmatched segment is a missing path and an update should be made.

- Strong confidence if $S_i > T_2$. We strongly believe that the unmatched segment is a missing path and an update should be made.

5 Experimentations and Evaluation

This section describes the results obtained for the matching process and the analysis of unmatched segments for updating purposes.

5.1 Datasets Description

The experimentations of our approach have been performed on two datasets: authoritative roads extracted from BDTOPO© and crowdsourced traces collected from different sport activities websites. The test area is in the Vosges Mountains (France), chosen for its interesting characteristics: small mountain mixing dense forest with different canopies and open areas.

BDTOPO© is a topographic database roughly having 1 m accuracy and produced by the French NMA. The road network theme contains different types of roads such as highways, national and local roads, paths for different type of use (pedestrian, tractors, bike etc.). Each feature road has a set of thematic attributes such as name, origin of the data, elevation, etc.

A total of 437 traces (9,773 km) have been downloaded from hikers and mountain bikers' websites (randoGPS, tracesGPS, visuGPS, and VTTour). Points composing the trace are theoretically described by 2D coordinates (WGS-84), timestamps, and elevation. However, in a total of 300,000 points, 106,206 points (36.3%) lack timestamps, and 6,580 points (2.2%) lack elevation. Regarding the traces, 157 of them (35.9%) have no timestamps at all. However, even if the timestamp for GPS points is sometime missing, there is always a global timestamp for the trace in the GPX file, which inform on the actuality of the trace, even if this may reflect only the date of upload of the trace, and note the date of the trace.

5.2 Data Matching Results

Hereon, some results obtained by the matching process described in Sect. 3.1 are illustrated. For validation purposes a 'ground truth' is defined by carrying out an interactive data matching. A total number of 41 traces (920 km) are manually matched by three experts based on different sources of data (e.g. orthophoto, maps).

The thresholds used are empirically defined by analysing the data with respect with precision and recall. Thus the optimal thresholds for buffer size are: $T_{HA} = 20$ m and $T_{LA} = 40$ m.

Table 1 Matching results in the test area

Data matching	R (%)	P (%)	F ₁ -score
Candidate selection	68	68	0.68
Direction criterion	74	75	0.74
Neighbourhood criterion	77	80	0.78
Length criterion	77	84	0.80

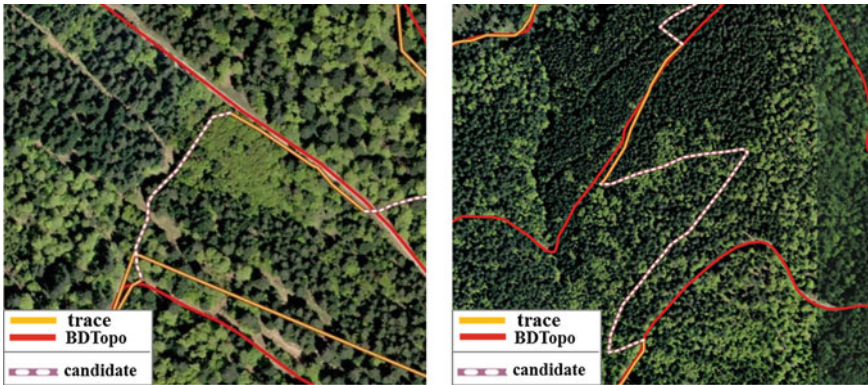


Fig. 4 Examples of efficiently unmatched segments

The threshold for the Direction Criterion, T_D , is empirically set to 30° to be not restrictive. For defining length criterion threshold, we follow the BDTOPO® specifications (BDTOPO® Specification 2.2, April 2017), which specify the minimum length for an independent path in authoritative roads. Thus, for the length criterion the T_L is set to 100 m. Finally, the same threshold of 100 m is defined for neighbourhood criterion to identify short segments.

Table 1 shows the importance of the criteria for the matching results. Thus, after the candidate selection, both precision and recall are poor, being equal to 68%. The direction criterion improved recall and precision for 6% and 7% respectively. After applying the neighbourhood criterion, recall and precision rose for respectively 3% and 5%. Finally, the length criterion improves only the precision of the unmatched matching results. Thus, each criterion improves the results, the most efficient one being the direction criterion.

Figure 4 illustrates two examples of a successful matching in detecting unmatched segments being candidate for missing paths for the next step (decision making).

However, some false negative and positive matching results are noticed. Figure 5 (on the left) illustrates a false positive matching where a segment is wrongly unmatched. Figure 5. (on the right) illustrates a false negative matching where a segment (in yellow) is wrongly matched to an authoritative road.

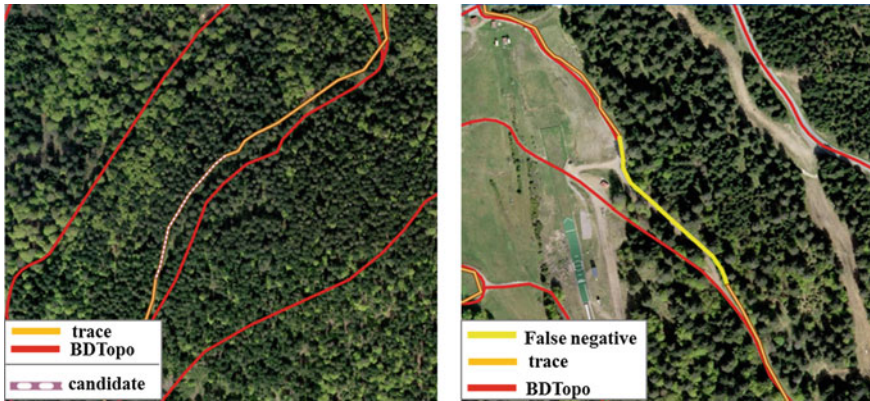


Fig. 5 Examples of false positive matching (on the left) and false negative matching (on the right)

5.3 Decision Making Results

The parameters necessary for the decision criteria are defined as follows. First, the parameters used in determining candidates following the same missing path criterion are: 80% of the length within the buffer, and the maximum angle between the candidates for same missing path is 30° . The threshold estimating the ‘nearest homologous path’, T_d , is empirically set to 30 m.

In total, 727.1 km of missing paths were found in the test area, which correspond to 7.7% of the total length of used traces. Compared to the total length of BDTOPO® road network in the test area, detected missing paths represent a contribution of 2.4%.

Concerning the combination of criteria to compute the degree of confidence for each identified missing path, at this stage of our work, it is difficult to define the priorities of criteria and weights, which needs to be refined according to NMA choices on the good compromise between number of detected potential updates and efforts necessary to examine each potential update. Thus, as suggested by Roszkowska (2013) the Equal Weight method (EW) is used and weights are represented as a uniform distribution on the unit n . Figure 6 illustrates the distribution of degree of confidence function of the percentage of missing paths. It can be noticed that relatively small amount of missing paths have a strong confidence. This is mainly due to the low actuality of our traces (e.g. from 2013 to 2015) which implies that all missing paths have weaker scores for the criterion actuality. According to the estimated thresholds (see Fig. 6) and the qualitative classification proposed in Sect. 4.1, 19% of missing paths has a weak confidence, 66% have a medium confidence, and 15% of missing paths have a strong confidence.

More importantly, a visual analysis of results allows us to classify the detected missing paths according to two criteria: configuration of missing paths related to the existing road network and interest of updating the missing path in the authoritative data. Let us mention that the visual analysis is carried out only for traces where

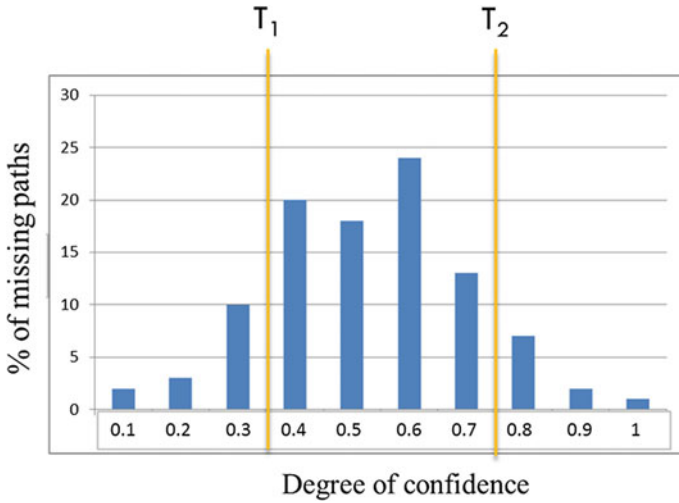


Fig. 6 Distribution of degrees of confidence for the missing paths

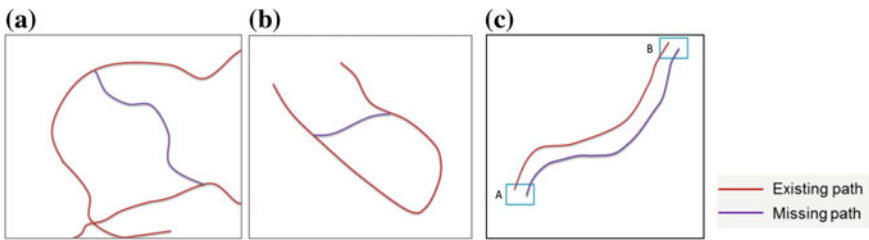


Fig. 7 Typical configurations of detected missing paths: **a** long independent paths; **b** shortcuts; **c** parallel paths

the ground truth has been defined. Concerning the first case, our analysis takes into account the main characteristics of missing paths regarding their functionality and relations with other paths from the road network. Three typical cases are identified: long independent paths, shortcuts, parallel paths and shown in Fig. 7.

The long independent paths type (illustrated in Fig. 7 on the left) is characterized by a significant length and an independent use from other paths in the network. This type of paths represents 42% (51.8 km) of the total missing paths. The second type is ‘shortcut’, as illustrated Fig. 7 (in the middle). Shortcuts have complementary use in the road network and usually save time and distance for navigation purposes. From the total amount of missing paths, shortcuts represent 26% (12.1 km). Finally, the last category, named ‘parallel road’ (Fig. 7, in the right) is characterised by a part of trace parallel from an existing authoritative path for a long distance, having the same role as authoritative path, for example connecting the same places, A and B. From the total amount of missing paths, parallel paths represent 32% (22.4 km). Figure 8 illustrates examples of detected missing paths from the test area.

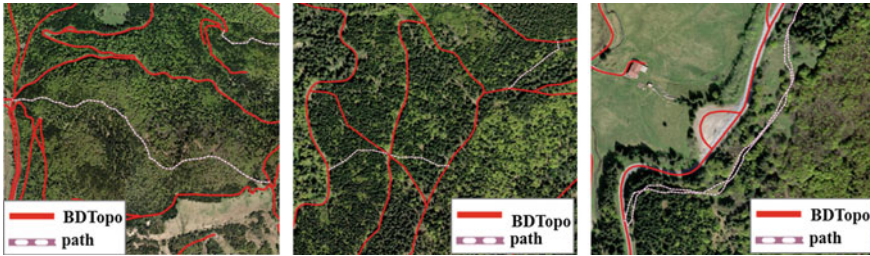


Fig. 8 Examples of detected missing paths from: long independent paths (on the left); shortcuts (on the middle); parallel paths (on the right)

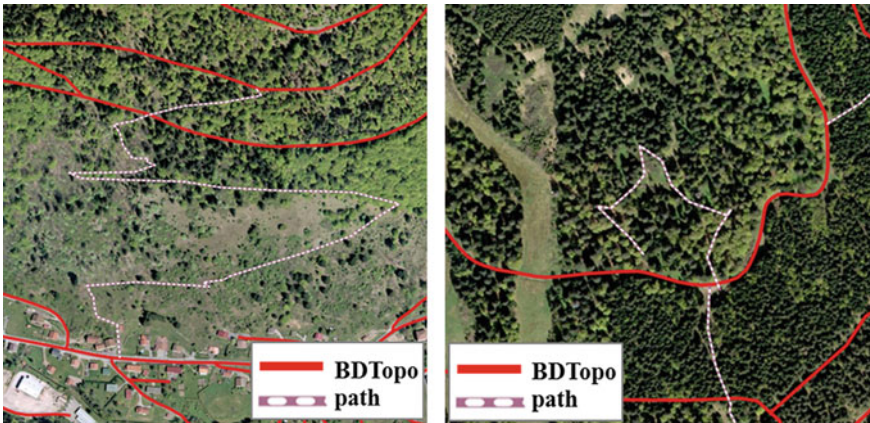


Fig. 9 Missing paths: doubtful missing path respecting data specification (on the left); path out of authoritative data specification (on the right)

Overall, the most represented missing paths are long independent paths, which have the highest priority for NMAs. With slightly more than one quarter, shortcuts are not negligible in missing paths. This can be explained mostly by human behaviour i.e. saving time and distance in their mobility by usually taking shortcuts when it is possible. Parallel roads, even representing almost one third of the missing paths, are with less interests for NMAs having the same role as the existing paths. Thus, according to the database specifications, they are not always necessary being considered.

Figure 9 (on the left) represents detected paths compliant with data specification but having a weak degree of confidence, whereas Fig. 9 (on the right) shows a detected path out of data specification and having a weak degree of confidence, and certainly an out of path behaviour. These examples show the necessity of checking the detected paths having small degree of confidence.

6 Conclusion

This research deals with studying the potential of using crowdsourced traces to detect updates in authoritative geographic data. We propose an approach based on two steps: matching data and decision making.

Data matching results consists in defining links with different cardinalities between homologous features. In this paper, we have focused on cardinality 1:0 (a feature crowdsourced data has not homologous in authoritative data) which may be interpreted such as ‘the missing feature should be added in authoritative data’. In future works, it would be interested to study the other types of cardinalities. For example, a cardinality 1:1 or n:m can be used as validation that the matched features still exist in the real world because are still in use. This is very important, especially for mountains paths, which frequently change. The case 0:1 can be interpreted such as ‘the feature does not exist anymore in the real world’ and should be erased from the authoritative data. This case requires more information and better quality and quantity of crowdsourced traces. This definitely requires other external sources for verification like satellite or orthophoto images, Flickr data, textual itinerary descriptions, etc. and integrating those heterogeneous data is challenge for future works.

Concerning decision making, some improvements could be made. First, the lack of metadata limits us in considering more information than only regarding the quantity of traces. DOP information is often used in the literature to evaluate accuracy in forest area. In our case, this information is not taken into account since most of the GPS points do not have DOP. Nevertheless, when DOP exists, it should be used and modelled such a new criterion for the data matching step. Other information such as traces’ owner is important. Chances that a new road has appeared are higher if traces are collected by different contributors. Moreover, time when the traces are collected is also important. If majority of traces are collected during a short time period, for example during the same day, the clue of new road creation is less reliable than if traces are collected in different periods of time (e.g. different days during a month). Second, in the current research the criteria have the same weights. The proposed four criteria are very different in terms of their nature and it is difficult to determine their relevance on the final decision without expert’s knowledge regarding their influences on final result. Therefore, one perspective is to define the weights of criteria by using objective information such as proposed in Chehreghan and Ali Abbaspour (2017). Concerning data sources, we plan to explore more crowdsourced data coming from other websites or OpenStreetMap (OSM). The last should be studied carefully due to the lack of completeness of data in mountain areas and to the fact that in mountain area we have noticed that data are bulk imports from sources such websites.

More globally, more research is needed to integrate updates coming from crowdsourced data into authoritative and should be done by taking into account the data production policies. Our results show that some detected single paths have high accuracy and may be integrated directly in authoritative data. Even with good accuracy, potential bias in path detection can be introduced because of use of traces coming from sport activities, when contributors may have defined their own paths. If many

paths are candidates to a single missing road (redundancy > 1; strong confidence) than a single geometry should be estimated. In this case, different methods proposed in the literature can be applied such as Etienne et al. (2015).

References

- Al-Bakri M, Fairbairn D (2012) Assessing similarity matching for possible integration of feature classifications of geospatial data from official and informal sources. *Int J Geogr Inf Sci* 26(8):1437–1456
- Bergman C, Oksanen J (2016) Conflation of OpenStreetMap and mobile sports tracking data for automatic bicycle routing. *Trans GIS* 20(6):848–868
- Bruns A (2009) *Blogs, Wikipedia, second life, and beyond*. Peter Lang, New York, N.Y
- Chehreghan A, Ali Abbaspour R (2017) An assessment of spatial similarity degree between polylines on multi-scale, multi-source maps. *Geocarto Int* 32(5):471–487
- Etienne L, Devogele T, Buchin M, McArdle G (2015) Trajectory box plot: a new pattern to summarize movements. *Int J Geogr Inf Sci* 30(5):835–853
- Fan H, Zipf A, Fu Q, Neis P (2014) Quality assessment for building footprints data on OpenStreetMap. *Int J Geogr Inf Sci* 28(4):700–719
- Girres J-F, Touya G (2010) Quality assessment of the French OpenStreetMap dataset. *Trans GIS* 14(4):435–459
- Ivanovic S, Olteanu-Raimond AM, Mustière S, Devogele T (2015) Detection of potential updates of authoritative spatial databases by fusion of volunteered geographical information from different sources. In: FOSS4G-Europe conference
- Jokar Arsanjani J, Helbich M, Kainz W, Darvishi Boloorani A (2013) Integration of logistic regression, Markov chain and cellular automata models to simulate urban expansion. *Int J Appl Earth Obs Geoinf* 21:265–275
- Koukoletsos T, Haklay M, Ellul C (2012) Assessing data completeness of VGI through an automated matching procedure for linear data. *Trans GIS* 16(4):477–498
- Lou Y, Zhang C, Zheng Y, Xie X, Wang W, Huang Y (2009) Map-matching for low-sampling-rate GPS trajectories. In: *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems—GIS '09*
- Liu C, Xiong L, Hu X, Shan J (2015) A progressive buffering method for road map update using OpenStreetMap data. *ISPRS Int J Geo-Inf* 4(3):1246–1264. <https://doi.org/10.3390/ijgi4031246>
- Mustière S, Devogele T (2008) matching networks with different levels of detail. *GeoInformatica* 12(4):435–453
- Newson P, Krumm J (2009) Hidden Markov map matching through noise and sparseness. In: *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems—GIS '09*. ACM Press, New York, NY, USA, pp 336–343
- Olteanu-Raimond A, Jolivet L, Van Damme M, Royer T, Fraval L, See L, Sturm T, Kärner M, Moorthy I, Fritz S (2018) An experimental framework for integrating citizen and community science into land cover, land use, and land change detection processes in a National Mapping Agency. *Land* 7(3):103
- Olteanu-Raimond A, Mustière S, Ruas A (2015) Knowledge formalization for vector data matching using belief theory. *J Spat Inf Sci* (10)
- Olteanu-Raimond A, Hart G, Foody G, Touya G, Kellenberger T, Demetriou D (2016) The scale of VGI in map production: a perspective on European National Mapping Agencies. *Trans GIS* 21(1):74–90
- Roszkowska E (2013) Rank ordering criteria weighting methods—a comparative overview. *Optim Stud Ekon* 5:14–33

- See L, Mooney P, Foody G, Bastin L, Comber A, Estima J, Fritz S, Kerle N, Jiang B, Laakso M, Liu H, Milčinski G, Nikšič M, Painho M, Pődör A, Olteanu-Raimond A, Rutzinger M (2016) Crowdsourcing, citizen science or volunteered geographic information? The current state of crowdsourced geographic information. *ISPRS Int J Geo-Inf* 5(5):55
- van Winden K, Biljecki F, van der Spek S (2016) Automatic update of road attributes by mining GPS tracks. *Trans GIS* 20(5):664–683
- Walter V, Fritsch D (1999) Matching spatial data sets: a statistical approach. *Int J Geogr Inf Sci* 13(5):445–473
- Xavier E, Ariza-López F, Ureña-Cámara M (2016) A survey of measures and methods for matching geospatial vector datasets. *ACM Comput Surv* 49(2):1–34
- Zielstra D, Hochmair H (2011) Comparative study of pedestrian accessibility to transit stations using free and proprietary network data. *Transp Res Rec J Transp Res Board* 2217:145–152

A Scalable Analytical Framework for Spatio-Temporal Analysis of Neighborhood Change: A Sequence Analysis Approach



Nikos Patias, Francisco Rowe and Stefano Cavazzi

Abstract Spatio-temporal changes reflect the complexity and evolution of demographic and socio-economic processes. Changes in the spatial distribution of population and consumer demand at urban and rural areas are expected to trigger changes in future housing and infrastructure needs. This paper presents a scalable analytical framework for understanding spatio-temporal population change, using a sequence analysis approach. This paper uses gridded cell Census data for Great Britain from 1971 to 2011 with 10-year intervals, creating neighborhood typologies for each Census year. These typologies are then used to analyze transitions of grid cells between different types of neighborhoods and define representative trajectories of neighborhood change. The results reveal seven prevalent trajectories of neighborhood change across Great Britain, identifying neighborhoods which have experienced stable, upward and downward pathways through the national socioeconomic hierarchy over the last four decades.

Keywords Neighborhood change · Sequence analysis · Spatio-temporal data analysis · Classification · Population dynamics

N. Patias (✉) · F. Rowe

Geographic Data Science Lab, Department of Geography and Planning, University of Liverpool, Liverpool, UK

e-mail: n.patias@liverpool.ac.uk

F. Rowe

e-mail: F.Rowe-Gonzalez@liverpool.ac.uk

S. Cavazzi

Ordnance Survey Limited, Southampton, UK

e-mail: stefano.cavazzi@os.uk

© Springer Nature Switzerland AG 2020

P. Kyriakidis et al. (eds.), *Geospatial Technologies for Local and Regional*

Development, Lecture Notes in Geoinformation and Cartography,

https://doi.org/10.1007/978-3-030-14745-7_13

1 Introduction

Changes over space and time reflect the complexity and evolution of demographic and socio-economic processes (Miller 2015). Yet measuring the magnitude, location and temporal frequency of these changes is challenging. Using traditional forms of data (i.e. census and survey data), demographic and socio-economic changes have often been captured at a very coarse temporal levels i.e. every month, year or decade. Also, this data is normally available at some spatially aggregated level. Administrative boundaries have traditionally been the default spatial framework for census and survey data collection and analysis (Goodchild 2013), and these areas are usually affected by boundary changes over time, particularly splitting an area in two (Casado-Díaz et al. 2017; Rowe et al. 2017a, b). So, a ‘freeze history’ approach has been generally employed to develop a consistent geography by freezing the zonal system at a certain point in time and systematically tracking subsequent alterations in geographical boundaries to amalgamate subsequently created areas (Rowe 2017).

Different levels of spatial aggregation can however produce different representations of a socio-economic process as a result of the Modifiable Areal Unit Problem (MAUP). MAUP refers to the statistical sensitivity and variability of results relating to the spatial framework of analysis (Openshaw 1983; Fotheringham and Wong 1991). The most appropriate spatial framework of analysis may thus differ according to the process in study (Prouse et al. 2014). MAUP can create ‘unreal’ spatial patterns which are caused by loss of information (Hayward and Parent 2009). Choosing areal units based on geographical coordinates, rather than aggregation of administrative boundaries, could help to tackle this issue by offering the possibility to analyze temporal data regardless of changes in geographic boundaries.

An increasing number of methods for spatio-temporal data analysis have been developed to study complex demographic and socio-economic processes, namely space-time point pattern, probabilistic time geography and latent trajectory models (An et al. 2015). Clustering techniques are often employed on space-time data, identifying patterns (Warren Liao 2005; Aghabozorgi et al. 2015; Arribas-Bel and Tranos 2018). There is also a wide variety of spatio-temporal statistical techniques in current literature where traditional deterministic trend models, stochastic trend models and stochastic residual models have been generalized to capture spatiotemporal processes using point-level data (Kyriakidis and Journel 1999), as well as spatial and temporal correlation using Spatio Temporal Autoregressive (STAR) and Bayesian hierarchical models on areal data (Huang 2017). Yet, these models are often restricted on specific situations namely particular data format or geometry types and are not flexible or adaptable to contribute in scalable space-time analysis frameworks. Spatio-temporal processes involve measurement of four dimensions namely occurrence, timing, order and duration of events or transitions and while the aforementioned methods can provide useful information about movements and points of interest, they only capture some dimensions of spatio-temporal processes. The integration of multiple approaches can provide context ‘aware’ data and expose patterns

based on analysis of the sequencing of events, rather than comparison of discrete points in time, capturing the full range of dimensions of spatio-temporal processes.

Sequence analysis provides a useful framework to integrate various analytical approaches and capture the four key dimensions of spatio-temporal processes i.e. occurrence, timing, order and duration of events or transitions. Sequence analysis was originally developed for analyzing DNA sequences (Sanger and Nicklen 1977; Bailey 2017), and theoretically introduced in the Social Sciences in the 1980s (Abbott 1983). Sequence analysis has recently been widely applied to analyze longitudinal individual family, migration and career trajectories (e.g. Rowe et al. 2017a, b; Backman et al. 2018).

Sequence analysis can also be applied to better understand the evolution of places. Conceptually, neighborhoods for example are assumed to progress through a number of pre-determined stages, transitioning through phases of development, growth, stability and decline (Hoover and Vernon 1959). However, prior empirical work has employed a static cross-sectional framework to explore these transitions between two points in time (e.g. Teernstra and Van Gent 2012) and assumed all neighborhoods undergo the same rigid pathway of change. These shortcomings partly reflect the lack of consistent spatial data over a longer window of time, but also the absence of an analytical approach to study these transitions in a temporally dynamic framework.

Only recently, empirical analyses have recognized the diversification in neighborhood transitions and enabled exploration and quantification of neighborhood change over a long period of time by using sequence analysis. Delmelle (2016) conducted a first study using sequence analysis for Chicago and Los Angeles, expanding her focus on a subsequent investigation to US 50 metropolitan areas over a 50 year period (Delmelle 2017). These studies contributed in providing a general approach for analyzing differentiating pathways of neighborhoods namely upgrading, downgrading or stable trajectories in the socio-economic hierarchy as well as gentrification processes. Yet they focused only on urban and metropolitan areas, missing the interaction between urban and rural continuum.

While these studies have advanced our understanding of neighborhood change in particular urban settings, significant gaps remain to be addressed. First, gridded data generation is needed to address the lack of consistent geographical boundaries over time (Janssen and Van Ham 2019). Second, the use of gridded data offers the potential to perform analyses at various geographical levels through aggregation of grids at particular administrative or functional areas. This opportunity provides a flexible dataset and a scalable approach for the use of purpose-built areas. Third, weighted clustering of the sequences provides a scalable approach on analyzing big datasets by separating the unique sequences matrix and their frequency in different vectors. This addresses the lack of information by using clustering approaches based on 'prototype' sequences where their frequency is not captured (i.e. how many neighborhoods followed the same sequence). Fourth, the aforementioned gaps are partly the result of the absence of a workflow that addresses the lack of temporally consistent geographical units and offers a way to effectively capture the key elements of changes in space and time (occurrence, timing, order and duration). This limitation

is addressed in this study by the integration of different approaches (i.e. population grid surface estimation, clustering analysis and optimal matching).

This paper aims to develop a scalable analytical framework for spatio-temporal data analysis addressing all four identified gaps. By doing so, it contributes to the current literature on spatio-temporal data analysis in three key ways:

1. By providing geographical consistent gridded data over a 40-year period for Great Britain;
2. By developing a scalable analytical framework in two ways: (i) offering a flexible dataset which can be aggregated at various geographical levels; and (ii) employing a weighted clustering approach to measure dissimilarity between individual sequences;
3. By formulating a workflow to effectively capture the key elements of changes in demographic and socio-economic process across space and over time through the integration of multiple approaches.

The remainder of this article is organized as follows. Section two describes the dataset and methods used in this research project, followed by results and discussion that are presented in section three. Finally, section four provides some concluding remarks and suggestions for further research.

2 Data and Methods

2.1 Data

The original data used in this study is drawn from five decennial Censuses for Great Britain (i.e. England, Scotland and Wales) covering the period from 1971 to 2011 with 10-year intervals. The five Censuses were conducted in 1971, 1981, 1991, 2001, 2011. The data was downloaded from the Office of National Statistics (ONS) portal (<http://casweb.ukdataservice.ac.uk> and <http://infuse.ukdataservice.ac.uk>).

Census administrative boundaries are not consistent over time. To this end, this paper uses an approach of recalculating Census counts from administrative boundaries to gridded data using *Popchange* project algorithm. *Popchange*, is a tool that provides population surfaces across Great Britain but also provides the algorithm which calculates correspondence between low-level Census administrative geographies and 1 km² grids (Lloyd et al. 2017). For this project, raw Census data covering a range of demographic, socio-economic and housing variables was downloaded in low-level Census administrative geographies (i.e. enumeration districts for 1971, 1981 and 1991 and output areas for 2001 and 2011) and *Popchange* algorithm used to convert Census counts to 1 km² grid counts.

These grid counts data correspond to estimates of census variables. As they are generated from a coarser level of geography, there is certain degree of uncertainty around these estimates. However, they offer two key advantages. Firstly, they provide

Table 1 Variables used in the analysis

Demographic	Socio-economic	Housing
Children: 0–14 years old	Managerial occupations	Own occupied housing
Young persons: 15–29 years old	Non-manual workers	Private rented housing
Middle aged adults: 30 to 44 years old	Manual and other workers	Council rented (social) housing
Older adults: 45–64 years old	Private mode	Vacancy rate
Retired: 65+ years old	Public transport	
Born in United Kingdom (UK) and Republic of Ireland (ROI)	Active mode	
Born in Europe	Other mode (i.e. other and work from home)	
Born in rest of the world	Unemployment rate	
Proportion of students		

a consistent level of geography to make comparisons of spatial units over a period of time. Secondly, they provide an effective tool to address the MAUP in a spatio-temporal context by providing a spatial framework based on geographical coordinates (i.e. 1 km² grids), rather than some arbitrary level of geographical aggregation. Grids can be aggregated to create purpose-built geographical systems depending on the process under analysis.

A drawback of grids is that administrative areas in rural and remote areas are often larger than a grid. Thus, population counts that are split between two or more grids in an administrative area, resulting in a small number of population counts per grid. In this study percentages of the variables were calculated by grid and given the small number of counts per grid, the accuracy of the variables' estimation is low. To overcome this issue, only grids which encompass multiple small areas were considered. To this end, the 1 km² grid layer overlaid over the 2011 census Output Area boundaries for Great Britain. The final output is 16,035 grid cells covering the whole Great Britain. The grid cells containing zero values can be removed to aid visualization and mapping of the data.

This study measures neighborhood change across three dimensions: demographic, socio-economic and housing. Table 1 lists the set of census variables used to capture these dimensions, all of which are measured as percentages for each grid cell i.e. the grid-specific population aged 0–14 over the grid-specific total population across all age groups.

There is variation in the number of categories across census years. For example, a greater number of categories is available for socio-economic status in the 2001 and 2011 Census compared to earlier years. So, data have been aggregated to broader categories which are consistent through time. Also, note that information on students was not available in 1971; nonetheless, it is considered as an important variable and is therefore included for the analysis.

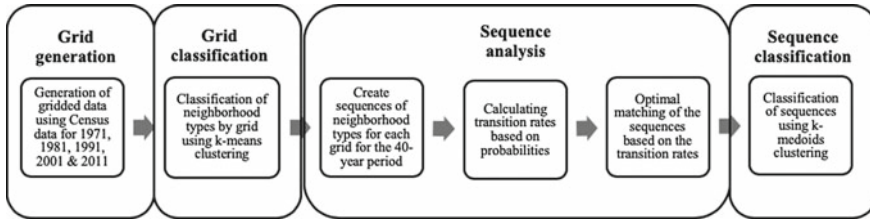


Fig. 1 Methodological framework workflow

2.2 Methods

The methodological framework developed in this study involves four main stages which can be divided into six steps, as presented in Fig. 1. In general, the first stage involves the production of gridded population data. This data is used in a second stage to create a geodemographic classification of neighborhoods based on the variables listed in Table 1, using k-means clustering. This classification provides representative types of neighborhoods. In a third stage, the classification is used to analyze the year-to-year transition of individual grids between neighborhood types and measure their similarity via optimal matching. In a final stage, this measure of similarity is employed to define a typology of representative neighborhood trajectories based on a k-medoids clustering. Details on each of these stages are provided next.

Stage 1. The official administrative boundaries used to collect the census data are not consistent over time. Boundary changes hamper temporal comparability of this data. Harmonization of these boundaries is needed to effectively track changes over time. To this end, *Popchange* algorithm was used to generate gridded data using raw data drawn from five decennial Censuses for Great Britain as described in Sect. 2.1.

Stage 2. Gridded population data is then used to create a geodemographic classification using a k-means algorithm. The input data is a pooled dataset of grids covering the whole Great Britain for all five census periods i.e. 80,175 grids (=16,035 grids * 5 years). A cluster analysis is performed on a pooled dataset including all five census periods to ensure consistency and comparability of cluster membership in the resulting partitioning solution. These are important elements for the longitudinal analysis of spatial data. For the k-means clustering algorithm, the number of k partitions, which define the number of cluster groups, need to be set prior to performing the analysis (Gentle et al. 1991). This has been set to eight performing 1,000 iterations. The approach used to specify the optimal number of clusters is a two-step sequential process. First, the sum of distances of each observation to their closest cluster center was calculated for a range of cluster options, from 3 to 15, creating an elbow curve. In an elbow curve, the sum of distances tends to decrease towards 0 as the k increases (the score is 0 when k is equal to the number of data points in the dataset, because then each data point belongs to its own cluster, with no error between the cluster and the center of the cluster). The goal is to determine the smallest number of k partitions that minimizes the sum of distances, and the elbow represents the point

at which diminishing returns by increasing k are achieved. Second, the k number at which these diminishing returns are achieved is used as the seed number of partitions. Various clustering partitions around this point were analyzed and mapped to determine the optimal number of clusters for this study i.e. eight. The output from the cluster analysis is temporally consistent geodemographic classification in which each year-specific grid cell is assigned to a neighborhood type.

Stage 3. This geodemographic classification is then used as input for sequence analysis. A key aim of this analysis is to define trajectories that characterize the ways in which the internal demographic and socio-economic structure of neighborhoods have changed over time. To this end, sequence analysis was used. Sequence analysis is built to analyze longitudinal categorical data and enables identification of representative patterns over a period of time. In the current study, the key aim is to identify a small number of representative trajectories of neighborhood change, and the application of sequence analysis involves three key steps. For the implementation of these steps, the TraMineR package in the R programming language was used (Gabadinho et al. 2009).

Step 1. The starting point is the creation of a sequence state object. A sequence state object refers to a dataset arranged in a wide format with rows identifying each spatial unit, columns identifying each time point, and individual cells indicating a specific state. To create a sequence state object, the geodemographic classification was used. Rows identify each geographical grid. Columns identify each of the five census years and each individual cell contains their corresponding year-specific neighborhood type. So, horizontally, each row provides a sequence of transition between different neighborhood type over the five census years.

Step 2. Sequences comparison requires a measure of the minimal cost of transforming one sequence to another. The operations can be used are insertion/deletion (i.e. indel) cost where a single value is specified to reflect how many insertions/deletions need to be made so that the two sequences match. But there is also the option of substitution cost matrix which is a square matrix of $s \times s$ dimensions, where s is the number of neighborhood types. So each (i, j) matrix element is the cost of substituting neighborhood type i with neighborhood type j . These elements called transition rates and are calculated based on the probability of transitioning from one neighborhood type to another. Then the optimal matching can be performed (i.e. measuring the similarity of those sequences) which is the sum of those rates for a given sequence.

Step 3. A key innovation of this study is the scalability of the developed framework to build and analyze sequences of neighborhood change. The calculation of dissimilarity between individual sequences is computationally intensive as it involves the use of substitution operations for each pair sequence in the dataset which increase proportionally with the number of spatial units and time points in the analysis.

The analysis of this paper involves the calculation of a dissimilarity matrix for 16,035 grids over 5 years; that is, a resulting matrix of 257,121,225 entries. In order to provide a scalable analytical framework, the unique sequences were identified and their frequencies were calculated and stored in different vectors. Then the unique (1,112) individual sequences used to compute the dissimilarity matrix of 1,236,544

entries. The idea behind this is that the dissimilarity matrix between all pairs of sequences has identical pairs (i.e. many grids that display the same transition, for example, from affluent to thriving neighborhoods). So, if only one pair is considered for the calculation and then it is expanded by the number of similar pairs in the dataset makes the computation less intensive. The use of the proposed approach can be applied to very large datasets for which the resulting dissimilarity matrix can go beyond the storage memory limits of R.

Timing of transitions between neighborhood types was considered a critical element for the definition of sequences as it helps discriminating between transitions resulting from structural economic changes and localized socio-economic shifts. To this end, substitution costs have been used capturing the temporal variation of transitions rather than indel costs which is static cost measure. The substitution cost between neighborhood types i and j for $i \neq j$ is computed by:

$$4 - p(i|j) - p(j|i) \quad (1)$$

where $p(i|j)$ is the transition rate between neighborhood types i and j . This probability is assumed to be dynamic reflecting the year to year transition between neighborhood types. So, a dynamic method of optimal matching was used which updates the substitution costs year to year to calculate distances between individual sequences. This method is referred as Dynamic Hamming method in the literature (Lesnard 2009).

Stage 4. The last stage involves producing a typology of neighborhood trajectories using the resulting sequence dissimilarity matrix from Stage 3. Partitioning Around Medoids (PAM) clustering method was selected for classifying sequences. It was preferred over hierarchical clustering methods because, although the PAM algorithm is similar to k-means, it is considered more robust than k-means as it can accept a dissimilarity matrix as an index and its goal is to minimize the sum of dissimilarities compared to k-means that it tries to minimize the sum of squared Euclidean distances (Gentle et al. 1991). The PAM is based on finding k representative objects or medoids among the observations and then k clusters (that should be defined as in k-means) are created to assign each observation to its nearest medoid.

As described in Stage 3 two vectors were created. One stores the dissimilarity matrix of the unique sequences and the other stores its sequence's frequency. The last issue that had to be tackled was the use of both vectors in a clustering algorithm, avoiding the creation of 'prototype' clusters but considering the whole dataset. In some hierarchical clustering methods (i.e. single linkage and complete linkage), the frequency of unique sequences does not affect the resulting partition of the data. But in the PAM algorithm, the number of observations in the matrix plays a role in the final result as it attempts to minimize the distance between each data point. Large datasets (e.g. of 47,000) result in a dissimilarity matrix of large dimensions (e.g. more than 2 billion), which cannot be handled in R where the storage memory limit is 2.1 billion.

An approach to overcome this problem is data weighting. This study applies a weighted version of the PAM clustering algorithm. The functionality of weighted

PAM clustering method is the same as using the usual PAM clustering but reduces the amount of memory needed to perform calculations over large datasets. To implement this method, a vector of the number of each unique sequence in the dataset was created and then used to weight dissimilarity matrix of these sequences when applying the PAM clustering method. In this way, the complete dataset (i.e. 16,035 sequences) was used in a scalable, faster and less computationally intense process. To implement this approach, ‘WeightedCluster’ R package was used (Studer 2013). For the Weighted version of PAM, the following function is minimized:

$$\text{minimize } \sum_{i=1}^n \sum_{j=1}^n w_{ij} d(i, j) z_{ij} \quad (2)$$

where d is the dissimilarity between each pair of sequences and z is a variable ensuring that only the dissimilarity between entities from the same cluster is computed. The term w is the weight parameter, which in this study is the frequency of each unique sequence. Consequently, in the weighted PAM method the dimensionality of the full dissimilarity matrix is reduced by creating a vector that contains the frequency of each unique sequence.

3 Results and Discussion

This section illustrates the results of Stages 2, 3 and 4 described in Sect. 2 in sequential order, starting by the geodemographic classification before discussing the sequence analysis and clustering of sequences to create representative neighborhood trajectories.

3.1 Temporal Clustering

As described in Sect. 2, k-means clustering was performed to create a geodemographic classification of neighborhoods, considering a 40-year period from 1971 to 2011 for Great Britain. Twenty-one variables were included in the analysis, covering demographic, socio-economic and housing characteristics. Eight clusters were returned as the optimal solution.

Figure 2 illustrates the mean variable values for each cluster. The name and key features of the eight neighborhood types are described below and displayed in Fig. 3:

- **Affluent:** These are the most affluent areas with most of the population belonging to the managerial socio-economic group with high proportion of population from abroad (10%). These areas are usually suburban and their populations mainly travel to work via private cars (53%). Public transport mode to work is used by

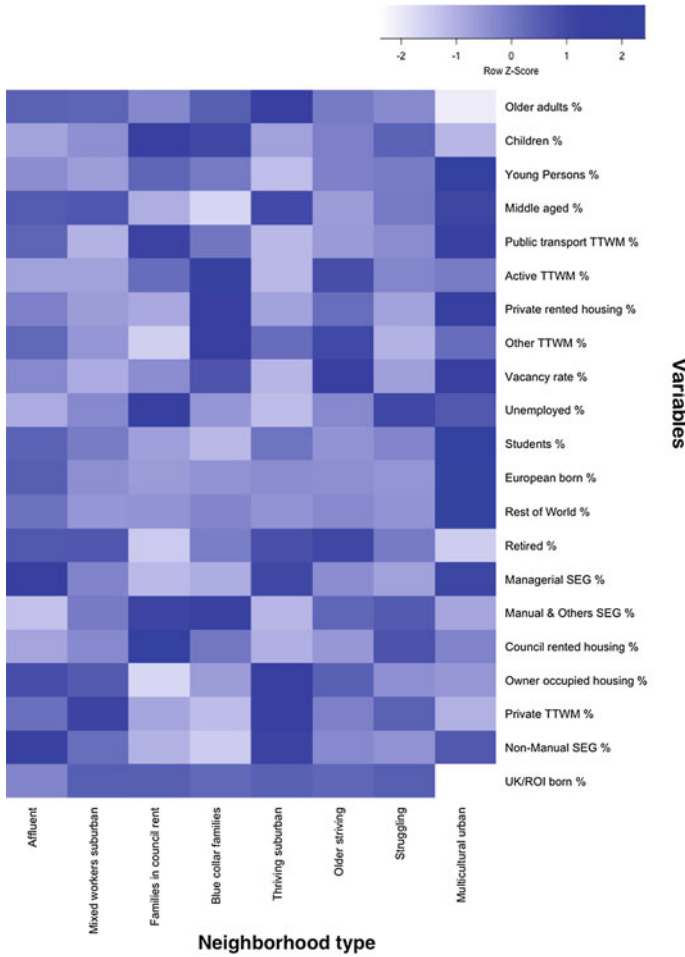


Fig. 2 Representative variables across neighborhood type

- 25% reflecting good public transport connections to workplace areas. These areas also have a high proportion of students (4.5%) and owner-occupied houses (76%).
- Mixed workers suburban: This group of neighborhoods is characterized by a mixture of people in manual (46%) and non-manual (43%) socio-economic groups with only a few students (3%). Their residents are largely UK and Republic of Ireland born (96%). There is high proportion of people travel to work with private mode of transport (70%) and finally high proportion of owner-occupied housing (70%).
 - Families in council rent: These neighborhoods are predominantly occupied by UK and Republic of Ireland born people (96%). There is high unemployment rate (11%), with high proportion of people staying in council rented housing (77%).

Finally, well connected or close to workplace areas as people use public (36%) and active mode of travel to work (22%).

- **Blue collar families:** These areas characterized by high proportion of manual workers (66%) owing a house (41%) that are predominantly UK and Republic of Ireland born (94%). Close to workplace areas with high proportion of people using active mode to travel to work (38%). This cluster appears in the earlier census years.
- **Thriving suburban:** These neighborhoods are quite similar to the Affluent areas with the difference of less people belonging to managerial socio-economic Group (18%) and higher ratio of owner-occupied houses (87%). Mainly using private mode to travel to work (74%) and low vacancy rate (4%) which shows that the demand for housing is high.
- **Older striving:** These neighborhoods are occupied by older people. Mainly manual workers (52%) but with few non-manual (38%) occupations too. There is high vacancy rate (7%) which represents low demand and thus people can afford to buy properties in these areas. The name of the cluster is Older striving but there are people from higher socio-economic Groups (i.e. non-manual and managerial occupations) living in these areas due to the affordability of housing.
- **Struggling:** Young and middle-aged families UK and Republic of Ireland born (96%) with high unemployment rate (10%) and an even split of people living in council rented (47%) or owner-occupied housing (46%). These neighbourhoods consist of -mainly- manual workers (56%) with few people in non-manual (37%) occupations.
- **Multicultural urban:** The two main characteristics of these neighborhoods are the high proportion of young people (29%) and high ratio of people born abroad (30%), which makes them highly ethnically diverse. There is a mixture of socio-economic Groups and high ratio of people relying on public (40%) or private (34%) transport to travel to work. It is also worth mentioning the high vacancy rate (7%) of these locations which are predominantly in city centers of urban areas, not the most 'desired' locations for housing in Great Britain.

This classification can be used to analyze spatio-temporal changes of neighborhood types. For example, a marked decrease in the number of blue collar families and families in council rent can be observed across Great Britain over the 40-year period. Liverpool emerges a prominent example changing from predominantly pink and purple in 1971 to red and yellow in 2011 in Fig. 3. The number of multicultural urban neighborhoods have significantly increased from 1971, especially between 2001 and 2011. These changes at the neighborhood level reflect the shrinkage of manual jobs in Great Britain after 1970s and the ethnic diversification of urban centers in the 2001 and 2011.

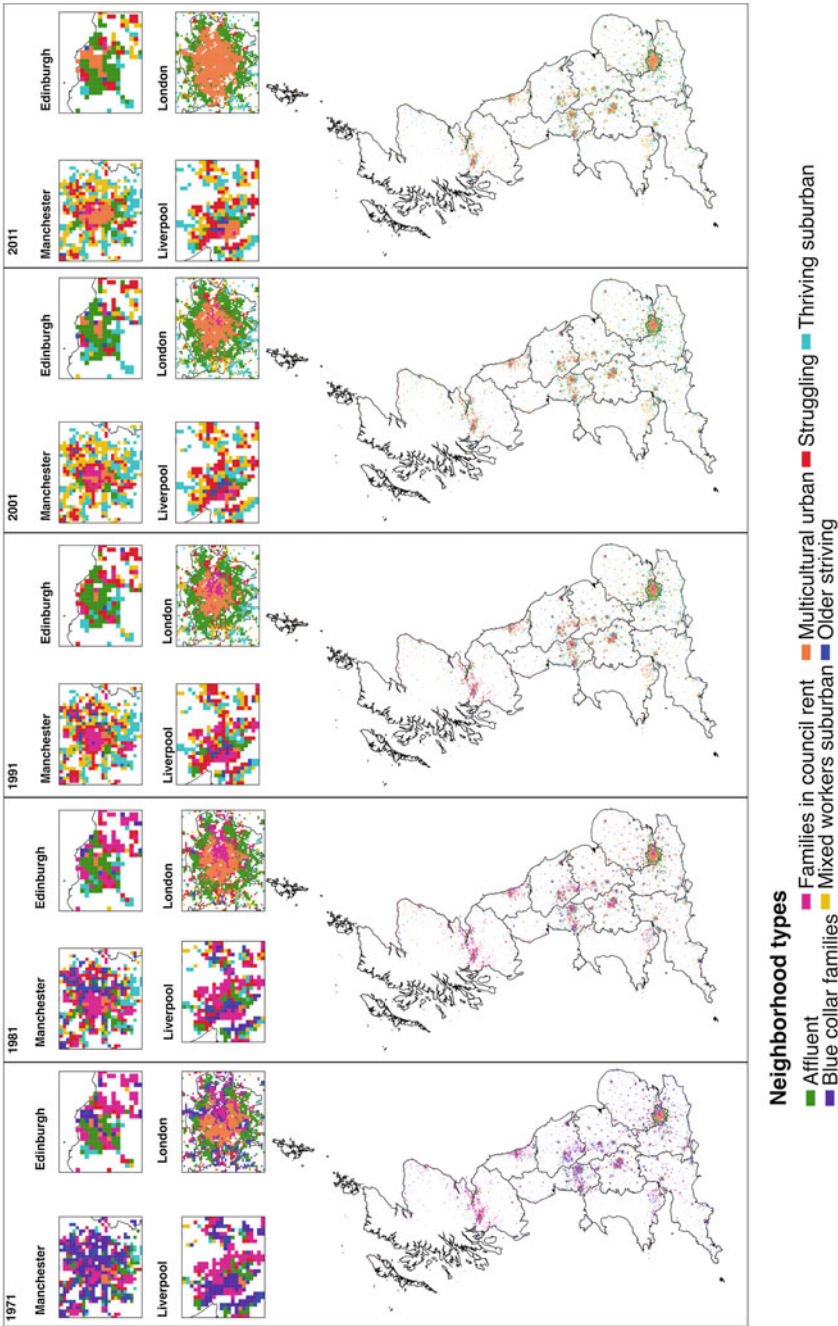
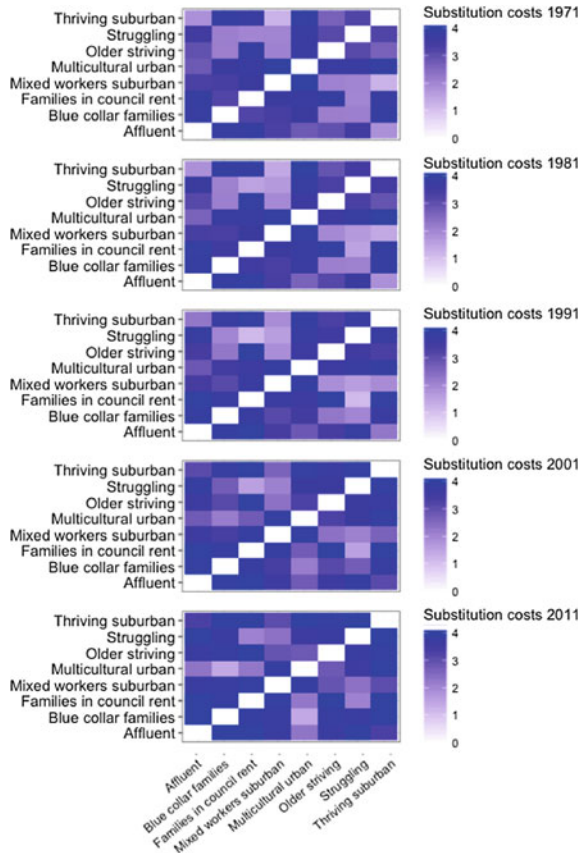


Fig. 3 Temporal neighborhood clusters in Great Britain

Fig. 4 Substitution costs from 1971 to 2011



3.2 Sequence Analysis

Using sequence analysis, a more comprehensive understanding of spatiotemporal process can be achieved by examining the occurrence, timing, duration and order of transitions between neighborhood types. Figure 4 displays the year-to-year substitution cost matrix to define the sequence of a neighborhood. It shows that lower substitution costs for earlier years, reflecting the higher degree of neighborhood transformation from 1971 to 2001. Neighborhoods during the 1970s were more likely to transition between mixed workers suburban to thriving suburban.

In addition to this, the results show that some neighborhood transitions between particular types are more common than others. Thus, the probability of transitioning between affluent and thriving suburban or blue collar families and struggling is higher compared to the probability of transitioning between affluent and blue collar families through all the decades. Yet from 2001 onwards these probabilities have been decreased as mentioned above.

3.3 Sequences Clustering

The substitution costs matrices were then used, to calculate a dissimilarity matrix between individual sequences and derive a typology of neighborhood sequences using the Dynamic Hamming method and weighted PAM clustering. Figure 5 displays the resulting typology of seven neighborhoods representing pathways of stability, improvement and decline across the national socio-economic hierarchy. Three panels of graphs are shown in Fig. 5. The top panel shows individual sequences. Each line in this graph represents a neighborhood. Each color denotes a particular type of neighborhood and the x-axis represents each census year. So, horizontally each line shows the transition of a neighborhood between neighborhood type over time. The middle panel displays the year-specific distribution of each neighborhood type. The bottom panel shows the mean time remaining in each neighborhood type.

The name and key features of the seven main neighborhood transition patterns are described below and displayed in Fig. 6:

- Stable affluent neighborhoods: Areas remaining persistently affluent over 1971 and 2011.
- Ageing manual labor neighborhoods: Areas transitioning from being dominated by blue collar families to an older striving neighborhood type.
- Increasingly socio-economically diverse neighborhoods: Areas transitioning from a struggling or blue collar families type to a mixed workers suburban type.
- Increasingly struggling home-owners neighborhoods: Areas transitioning from a families in council rent type to a struggling type.
- Stable multicultural urban neighborhoods: Areas remaining multicultural in urban locations.
- Rejuvenating neighborhoods: Areas transitioning from an older striving type to a mixed workers suburban type.
- Up-warding thriving neighborhoods: Areas transitioning from an older striving type to, or remaining in, a thriving suburban type.

The spatial distribution of these neighborhood trajectories varies between and within areas. There are areas such as Edinburgh and London suburbs that are dominated by stable affluent neighborhoods, while others such as Liverpool and Newcastle have more increasingly struggling home-owners neighborhoods. Regarding the distribution of neighborhood trajectories within areas there are few interesting patterns. One example is the rejuvenating neighborhoods that are characterized by younger people with various socioeconomic backgrounds ‘replacing’ the older population in suburban areas. Another example is the upward trajectories from struggling neighborhoods to more socio-economic diverse and the massive increase of thriving neighborhoods in suburban areas too. Lastly, stable multicultural urban areas appear in or close to city centers of all big urban conurbations.

Finally, ageing of British population is clearly reflected in the results. Suburban and rural areas are largely occupied by retired and older people. Interestingly, this pattern has changed slightly in the last decade, reflecting more inclusive communities both socio-economically and ethnically.

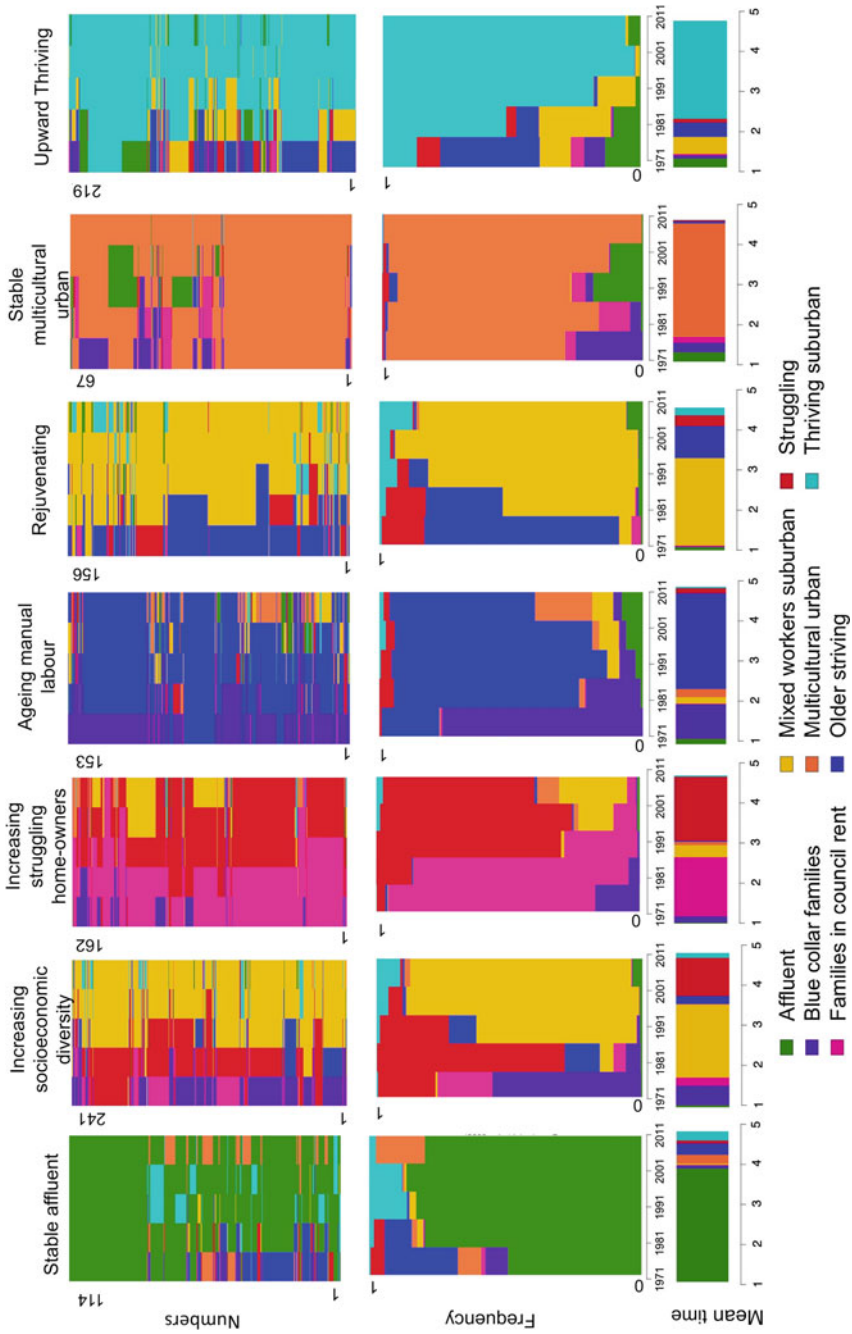


Fig. 5 Neighborhood trajectories clusters

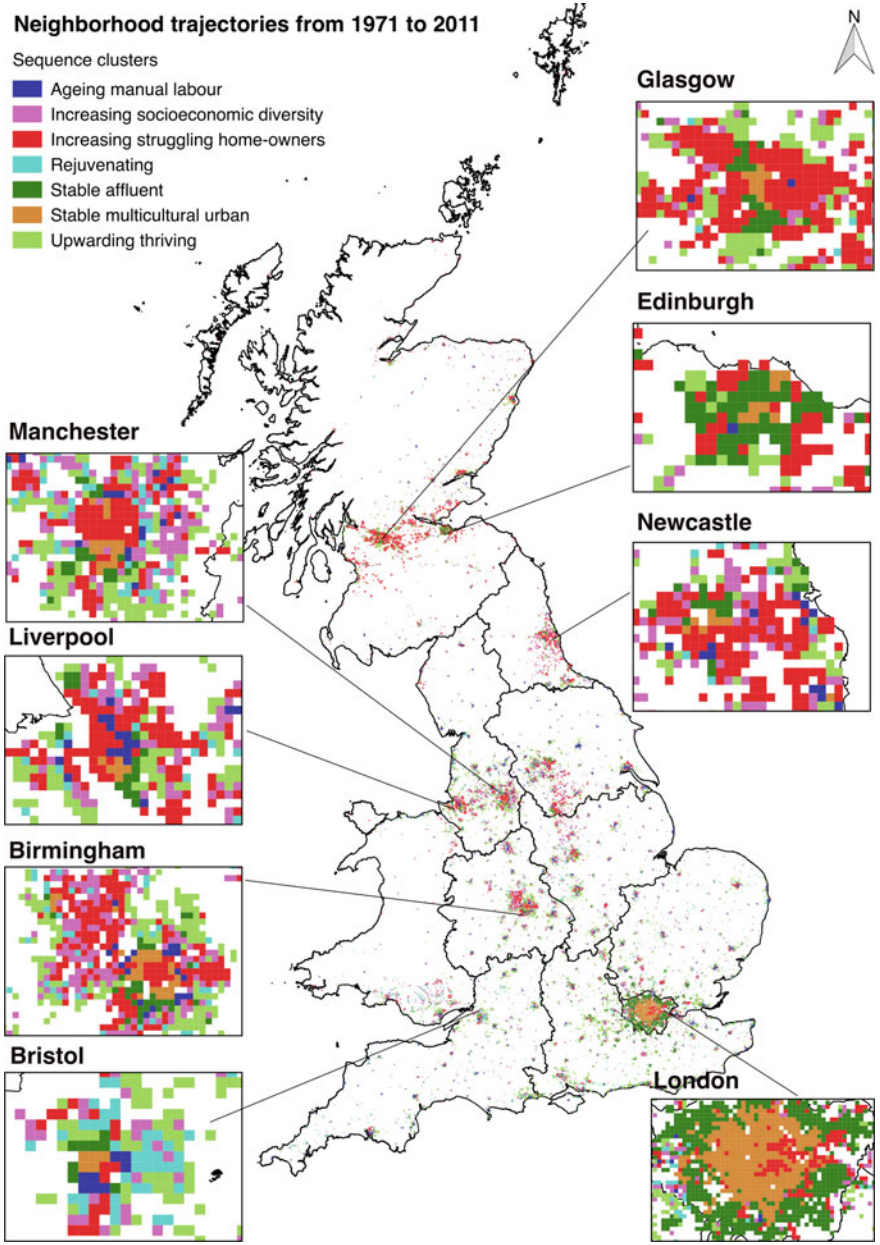


Fig. 6 Neighborhood trajectories map

4 Conclusion

This study proposed a novel scalable analytical framework for spatiotemporal data analysis. It does so by (1) producing a temporally consistent spatial framework and geodemographic classification based on 1 km² grids; (2) offering the potential to perform analysis at particular administrative, functional or purpose-built areas; (3) implementing a weighted approach to measure dissimilarity between individual neighborhood trajectories; and, (4) integrating multiple approaches (population grid surface estimation, clustering analysis and optimal matching) to analyze long-term change.

The proposed spatiotemporal analytical approach offers a framework within which the evolution of complex demographic and socio-economic processes can be effectively captured and enables understanding of the ways in which past conditions influence present and future transitional changes. Unlike commonly used longitudinal approaches such as event history analysis, which focuses on a single transition, the proposed sequence analysis provides a more comprehensive representation of present and future changes by examining the chronological sequence of events. Such approach enables to unravel key dimensions of changing socio-economic processes in terms of their incidence, prevalence, duration, timing and sequencing—which can serve as useful guidance for policy development.

The proposed approach offers the potential to expand understanding on key demographic and socio-economic processes. A key area for future research is the analysis of trajectories of socio-inequality examining at various levels of spatial aggregation and determining the extent of intra-regional and inter-regional inequalities. Such analysis can guide policy intervention by identifying spatial concentration of poverty and areas undergoing continuous economic decline. Another area of future investigation is the analysis of population change by identifying areas experiencing rapid and continuous population loss or population ageing in the light of sustained low patterns of fertility and signs of declining life expectancy (Green et al. 2017).

References

- Abbott A (1983) Sequences of social events: concepts and methods for the analysis of order in social processes. In: Abbott A (ed) *Historical methods*; Fall 1983; 16, 4; Periodicals Archive Online pg. 129'
- Aghabozorgi S, Seyed Shirkorshidi A, Ying Wah T (2015) Time-series clustering—a decade review. *Inf Syst* 53:16–38. <https://doi.org/10.1016/j.is.2015.04.007>
- An L et al (2015) Space–time analysis: concepts, quantitative methods, and future directions. *Ann Assoc Am Geogr* 105(5):891–914. <https://doi.org/10.1080/00045608.2015.1064510>
- Arribas-Bel D, Tranos E (2018) Characterizing the spatial structure(s) of cities “on the fly”: the space-time calendar. *Geogr Anal* 50(2):162–181. <https://doi.org/10.1111/gean.12137>
- Backman M, Lopez E, Rowe F (2018) Career trajectories and outcomes of forced migrants in Sweden: self-employment, employment or persistent inactivity? *Small Bus Econ*
- Bailey TL (2017) *Bioinformatics*. <https://doi.org/10.1007/978-1-4939-6622-6>

- Casado-Díaz JM, Martínez-Bernabéu L, Rowe F (2017) An evolutionary approach to the delimitation of labour market areas: an empirical application for Chile. *Spat Econ Anal* 12(4):379–403. <https://doi.org/10.1080/17421772.2017.1273541>
- Delmelle EC (2016) Mapping the DNA of urban neighborhoods: clustering longitudinal sequences of neighborhood socioeconomic change. *Ann Am Assoc Geogr* 106(1):36–56. <https://doi.org/10.1080/00045608.2015.1096188>
- Delmelle EC (2017) Differentiating pathways of neighborhood change in 50 U.S. metropolitan areas. *Environ Plan A* 49(10):2402–2424. <https://doi.org/10.1177/0308518X17722564>
- Fotheringham AS, Wong DWS (1991) The modifiable areal unit problem in multivariate statistical analysis. *Environ Plan A* 23(7):1025–1044. <https://doi.org/10.1068/a231025>
- Gabardinho A et al (2009) Mining sequence data in R with the TraMineR package: a user's guide for version 1.1.1, pp 1–129
- Gentle JE, Kaufman L, Rousseeuw PJ (1991) Finding groups in data: an introduction to cluster analysis. *Biometrics*. <https://doi.org/10.2307/2532178>
- Goodchild MF (2013) Prospects for a space-time GIS. *Ann Assoc Am Geogr* 103(5):1072–1077. <https://doi.org/10.1080/00045608.2013.792175>
- Green MA et al (2017) Could the rise in mortality rates since 2015 be explained by changes in the number of delayed discharges of NHS patients? *J Epidemiol Community Health* 71(11):1068–1071. <https://doi.org/10.1136/jech-2017-209403>
- Hayward P, Parent J (2009) Modeling the influence of the modifiable areal unit problem (MAUP) on poverty in Pennsylvania. *Pa Geogr* 47(1):120–135
- Hoover EM, Vernon R (1959) Anatomy of a metropolis; the changing distribution of people and jobs within the New York metropolitan region, New York metropolitan region study. Harvard University Press, Cambridge, MA (New York metropolitan region. Study: no. 1). <http://mirlyn.lib.umich.edu/Record/004478493>
- Huang B (2017) Comprehensive geographic information systems. Elsevier
- Janssen HJ, Van Ham M (2019) Resituating the local in cohesion and territorial development report on multi-scalar patterns of inequalities. <https://doi.org/10.13140/rg.2.2.23832.65287>
- Kyriakidis PC, Journel AG (1999) Geostatistical space-time models: a review. *Math Geol* 31(6):651–684. <https://doi.org/10.1023/A:1007528426688>
- Lesnard L (2009) Setting cost in optimal matching to uncover contemporaneous socio-temporal patterns. *Sociol Methods Res*. <https://doi.org/10.1177/0049124110362526>
- Lloyd CD et al (2017) Exploring the utility of grids for analysing long term population change. *Comput Environ Urban Syst* 66:1–12. <https://doi.org/10.1016/j.compenvurbsys.2017.07.003>
- Miller HJ (2015) Space-time data science for a speedy world. *I/S J Law Policy Inf Soc* 10(3):705–720. <https://doi.org/10.1525/sp.2007.54.1.23>
- Openshaw S (1983) The modifiable area unit problem. *Concepts Tech Mod Geogr* 38:1–41. <https://doi.org/10.1177/1077558707312501>
- Prouse V et al (2014) How and when scale matters: the modifiable areal unit problem and income inequality in Halifax. *Can J Urban Res* 23(1):61–82
- Rowe F (2017) The CHilean Internal Migration (CHIM) database: temporally consistent spatial data for the analysis of human mobility. *Region* 4(3):1. <https://doi.org/10.18335/region.v4i3.198>
- Rowe F, Casado-Díaz JM, Martínez-Bernabéu L (2017a) Functional labour market areas for Chile. *Region* 4(3):7. <https://doi.org/10.18335/region.v4i3.199>
- Rowe F, Corcoran J, Bell M (2017b) The returns to migration and human capital accumulation pathways: non-metropolitan youth in the school-to-work transition. *Ann Reg Sci* 59(3):819–845 (Springer, Berlin, Heidelberg). <https://doi.org/10.1007/s00168-016-0771-8>
- Sanger F, Nicklen S (1977) DNA sequencing with chain-terminating 74(12):5463–5467
- Studer M (2013) WeightedCluster library manual, pp 1–34. <https://doi.org/10.12682/lives.2296-1658.2013.24>

Teernstra AB, Van Gent WPC (2012) Puzzling patterns in neighborhood change: upgrading and downgrading in highly regulated urban housing markets. *Urban Geogr* 33(1):91–119. <https://doi.org/10.2747/0272-3638.33.1.91>

Warren Liao T (2005) Clustering of time series data—a survey. *Pattern Recogn* 38(11):1857–1874. <https://doi.org/10.1016/j.patcog.2005.01.025>

Improving Business-as-Usual Scenarios in Land Change Modelling by Extending the Calibration Period and Integrating Demographic Data



Romain Mejean, Martin Paegelow, Mehdi Saqalli and Doryan Kaced

Abstract Land use and land cover change (LUCC) models are increasingly being used to anticipate the future of territories, particularly through the prospective scenario method. In the case of so-called trend or Business-as-Usual (BAU) scenarios, the aim is to observe the current dynamics and to extend them into the future. However, as they are implemented as baseline simulation in most current software packages, BAU scenarios are calibrated from a training period built from only two dates. We argue that this limits the quantitative estimation of future change intensity, and we illustrate it from a simple model of deforestation in Northern Ecuadorian Amazon using the Land Change Modeler (LCM) software package. This paper proposes a contribution to improve BAU scenarios calibration by mainly two enhancements: taking into account a longer calibration period for estimating change quantities and the integration of thematic data in change probabilities matrices. We thus demonstrate the need to exceed the linear construction of BAU scenarios as well as the need to integrate thematic and particularly socio-demographic data into the estimation of future quantities of change. The spatial aspects of our quantitative adjustments are discussed and tend to show that improvements in the quantitative aspects should not be dissociated from an improvement in the spatial allocation of changes, which may lead to a decrease in the predictive accuracy of the simulations.

Keywords Land change modelling · Business as usual · Scenarios · Training dates

R. Mejean (✉) · M. Paegelow
GEODE UMR 5602 CNRS, Université Toulouse 2 Jean Jaurès, Toulouse, France
e-mail: romain.mejean@univ-tlse2.fr

M. Paegelow
e-mail: paegelow@univ-tlse2.fr

M. Saqalli
GEODE UMR 5602 CNRS, Toulouse, France
e-mail: mehdi.saqalli@univ-tlse2.fr

D. Kaced
IRIT UMR 5505, CNRS, Université Toulouse 1 Capitole, Toulouse, France
e-mail: doryan.kaced@irit.fr

1 Introduction

Over the past several decades, geographers developed a large spectrum of models to study land systems through land change, also called land use and land cover change (LUCC), whose socio-environmental impacts have been demonstrated (Chhabra et al. 2006; Mahmood et al. 2014; Oliver and Morecroft 2014). Among them, pattern-based models (PBM) of LUCC (Camacho Olmedo et al. 2018) are spatially explicit models allowed by knowledge of the drivers of change (Lambin et al. 2001; Carr 2003) and by change analysis methods (Mas 1999; Lambin et al. 2001; Comber et al. 2016). The purpose of PBM is to anticipate future changes in order to guide the present action e.g. in terms of public policy, by using prospective scenarios (Houet and Gourmelon 2014). Thus, the prospective scenario technique can be used both to observe the continuation of past and current trends in the future and to project alternative pathways (Veldkamp and Lambin 2001; Escobar et al. 2018). We will focus here on the first approach, called “business as usual” (BAU) scenarios which is a path-dependent approach (Houet et al. 2016) consisting therefore in extending the trend observed in the past over time. BAU scenarios are frequently found in the literature on LUCC modelling (Escobar et al. 2018) as well as in many cases of PBM application, in particular because PBM software packages includes BAU scenarios as baseline simulation (Mas et al. 2014).

According to Mas et al. (2014), the modelling process implemented in these PBM software packages can be divided into five steps: quantity of change estimate, change potential evaluation, spatial allocation of change, reproduction of temporal and spatial patterns and model evaluation. Although there is extensive literature on improving the spatial allocation of changes or model evaluation (Pontius and Millones 2011; Maestriperi and Paegelow 2013), there is little work on improving the quantitative estimation of change intensity. Indeed, most of the time, applying a BAU scenario means defining a single calibration period, between two training dates, according to the available data (Mas et al. 2018). The model uses notably this calibration period to estimate future change quantities, using generally one-order Markov chains (Camacho Olmedo and Mas 2018). Indeed, present-time software only allow the use of only two training dates (e.g. Land Change Modeler, CA_Markov, Dinamica EGO, Metronamica, ApoLUS, LucSim) and it has been shown that the choice of training period is not insignificant and that the simulation results obtained are different according to the training dates that have been chosen (Paegelow et al. 2014; Paegelow 2018).

The spatial expansion of the agricultural frontier in Northern Ecuadorian Amazon (NEA) can be observed over time from historical remote sensing images: settlement patterns and the forest clearing they induce are identifiable by their familiar fish-bone patterns, spread alongside the roads (Baynard et al. 2013). In the NEA, Mena et al. (2006) calculated an annual deforestation rate of 2.49% between 1986 and 1996 and of 1.78% between 1996 and 2002, i.e. a slowing of deforestation over time. We argue that, in a path-dependent approach like that of the BAU scenarios, such a slowdown in the rate of deforestation (i.e. in quantities of change) could not be deduced from

purely spatial and linear assumptions, e.g. from only two training dates, but rather requires taking into account a longer period of time and the consideration of thematic data.

Based on a simple model of deforestation dynamics in the NEA using the Land Change Modeler (LCM) software (Eastman and Toledano 2018), we propose here a contribution to improve BAU scenarios and more specifically the quantitative estimation of change intensity by mainly two enhancements. First, this contribution tries to exceed the linearity of BAU scenarios resulting from taking into account only two training dates. Then, authors introduce available, socio-economic, especially demographic, driver data directly to make more realistic classic Markov matrices. Both approaches are implemented by adjusting Markovian transition probabilities.

2 Materials and Methods

2.1 Context and Study Area

Northern Ecuadorian Amazon (NEA) is a region located in the western part of the Amazon basin, in the eastern part of Ecuador's national territory called "*Oriente*". This region is characterized by significant endemism to such a degree that it is known as one of the world's biodiversity hotspots (Orme et al. 2005). However, since the discovery of oil fields in the late 1960s, this territory is undergoing significant deforestation coupled with a fast population growth, due to free land accessibility, a high fertility rate and to a continue in-migration (Bilsborrow et al. 2004). Indeed, the road infrastructures built for oil extraction have enabled an agricultural colonization, mainly by small farmers from Andean and Coastal regions of Ecuador (Hiraoka and Yamamoto 1980; Bromley 1981; Brown et al. 1992). This agricultural colonization was spontaneous but also supported by the Ecuadorian authorities through two land reforms (Wasserstrom and Southgate 2013) and logistic support (Juteau-Martineau et al. 2014).

We will focus here on an area composed of a set of sub-watersheds, altogether surrounding and including the *parroquia* of Dayuma, inherited from another modelling approach dealing with environmental contamination (Houssou 2016). This area (Fig. 1), is located south of the city of Coca and Río Napo, in NEA. We have developed land cover classification for this study area using the following procedure detailed below.

2.2 Data and Image Processing

The land cover data used in the modelling process were obtained by using relatively simple image processing, coupling supervised segmentation (Paegelow and Camacho Olmedo 2010) and classification based on *maximum likelihood algorithm*. Then,

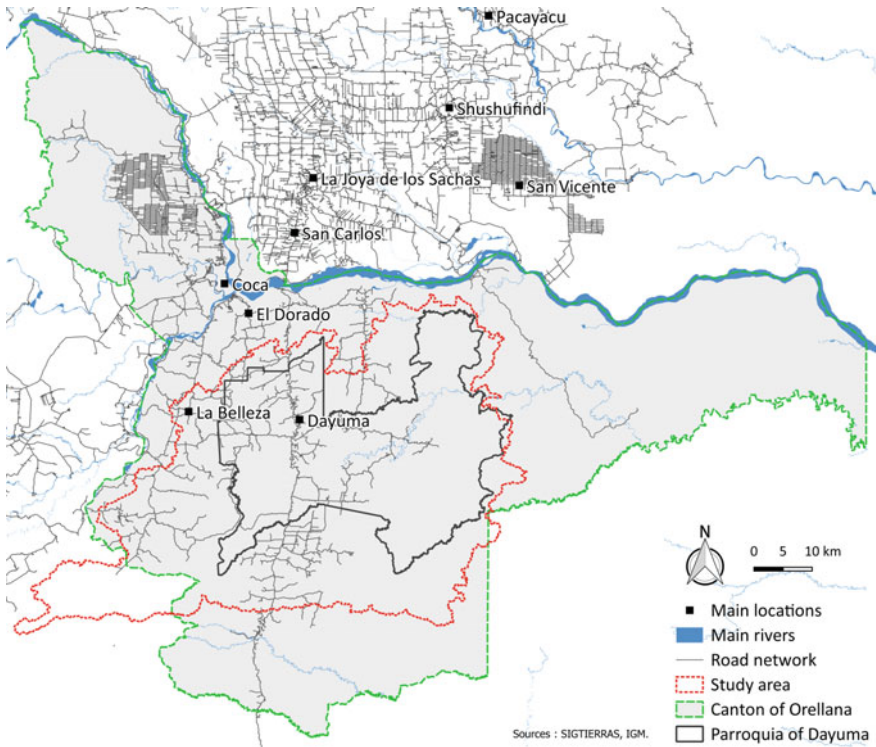


Fig. 1 Study area, in NEA

authors calculated the annual deforestation rates. In order to minimize classifications errors, we have chosen to classify land use into four major categories: water, forested areas, deforested areas (merging of the classes “bare soil”, “crops” and “pastures”) and urban areas. Tables 1, 2 and 3 indicate image characteristics, areas of each of the land cover classes and estimates of the annual rate of deforestation we have derived from it.

Despite some problems in detecting water surfaces, we observe a trend similar to that observed by Mena et al. (2006) at a different period and further north, in an earlier colonized territory: a slowing down of the annual rate of deforestation. Indeed, according to our data, while 0.61% of forests disappeared each year between 1998 and 2002 in our study area, only 0.29% was disappearing each year over the 2013–2017 period.

Table 1 Image characteristics

Satellite sensor	Path/row	Data acquired	Spatial resolution (m)
LANDSAT 5-TM	9/60 and 9/61	25 September, 1998	30 m
LANDSAT 7-ETM+	9/60 and 9/61	12 September, 2002	30 m
LANDSAT 8-OLI TIRS	9/60 and 9/61	2 September, 2013	30 m
SENTINEL-2		1 September, 2017	10 m decreased to 30 m by generalization ^a
SENTINEL-2		8 February, 2018	10 m decreased to 30 m by generalization ^a

^aReduction in the number of columns and rows while decreasing cell resolution by a pixel thinning algorithm

Table 2 Area per class (ha) on classification

Area per class (ha)	1998	2002	2013	2017
Water	53.46	0	5.67	0
Forested areas	259703.73	253322.01	240156.63	237400.83
Deforested areas	26889.39	33391.35	46470.15	49218.48
Urban areas	95.94	29.16	110.07	123.21

Table 3 Annual deforestation rate (%/year)

1998–2002	2002–2013	2013–2017
−0.61	−0.47	−0.29

2.3 Implemented Pattern-Based Model

The software package chosen for the pattern-based modelling process, called Land Change Modeler (LCM), is integrated into TerrSet (Eastman 2014) and is used to develop prospective models of LUCC, based on observations of past changes, statistical and machine learning methods to calibrate functions describing the relationship between change and drivers of change (Mas et al. 2018).

Although many authors have focused on analysing the functioning of LCM (Mas et al. 2014; Eastman and Toledano 2018), it is necessary to recall here some essential points about it as a brief overview: to estimate future change quantities, LCM uses Markov chains from a calibration period purposely defined by two training dates in order to determinate matrices of future transition probability between land use classes. LCM allows the use of an external transition probability matrix. Then, in terms of spatial allocation of changes, LCM allows the user to choose between three different methods to determinate the location of future changes, based on the relationships between driver variables loaded into the model and changes that occurred during the training period: (i) a multi-layer perceptron (MLP) neural network (Mas 2004), (ii) Similarity-Weighted Instance-based Machine Learning (SimWeight) and (iii) Logistic Regression. The simulation results can be expressed in two forms:

Table 4 Markovian matrix of transition probabilities

	Water	Forested areas	Deforested areas	Urban areas
Water	0	0.3333	0.3333	0.3333
Forested areas	0	0.9657	0.0343	0
Deforested areas	0	0.0882	0.9106	0.0012
Urban areas	0	0	0.2143	0.7857

Reading: from row to column

- (a) a soft simulation, i.e. a map of projected potential for transitions, mapping the places most prone to change. It can then be validated by means of a Receiver Operating Characteristic (ROC) analysis (Pontius and Schneider 2001; Mas et al. 2013)
- (b) a hard simulation, i.e. a qualitative map of projected LUCC, which can be validated by pixel-by-pixel validation techniques (Chen and Pontius 2010).

LCM is also able to integrate dynamic drivers into the modelling process (recalculated at each time step of the simulation), such as land use or road network as well as incentives or constraints for change e.g. the presence of protected areas that reduce deforestation. Finally, especially for us, BAU-type trend scenarios are included in LCM as baseline simulations.

The model we developed with LCM was trained over the period 2002–2013 and we used it to do projections for the year 2017. For simplification purposes, the only transition considered by the model is the transition from forested to deforested areas. Tables 4 and 5 below show respectively the Markovian matrix of transition probabilities calculated by LCM and the spatial driver variables used by the MLP (method we have chosen to spatially allocate future changes).

Drivers that are not Euclidean distances (ED) to relevant features like roads, oil fields or already deforested areas are socio-economic drivers selected from the last population census (Instituto Nacional de Estadística y Censos, INEC, 2010). More specifically, these are maps obtained by spatial interpolation (TIN method, Floriani and Magillo 2009) of census detail file data processed with REDATAM (De Grande 2016), from *localidades dispersas* and *manzanas*, which are census basic point units of the census. First, we selected the drivers to be included in the model based on our readings on deforestation processes in NEA or other South American contexts. So, we selected the variables we assumed to refer to: household size (Morin 2015), position of household in the lifecycle (Perz and Walker 2002), good living conditions (pull effect, Mena et al. 2006) and province of origin (push effect, to identify settlers). In a second step, among these drivers, we arbitrarily selected those with a Cramer V calculated with the class “deforested spaces” greater than 0.2.

LCM provides elements of model skill from the training process, based on analyses of a set of validation pixels: at each iteration, the MLP generates predicted class membership for each of the validation pixels and reports an overall accuracy rate and a skill score. According to the TerrSet documentation, the skill score represents the

Table 5 Spatial driver variables implemented in the model

Driver variable	Source	Cramer V
Viviendas with one to two rooms	INEC, 2010	0.2777
Viviendas with three to five rooms	INEC, 2010	0.2307
Viviendas with six or more rooms	INEC, 2010	0.2271
People aged 0–14 years old	INEC, 2010	0.2543
People aged 65 years old and more	INEC, 2010	0.2419
People with 6 or more children	INEC, 2010	0.2272
Viviendas connected to the electricity network	INEC, 2010	0.237
Viviendas with WC facilities	INEC, 2010	0.2912
People born in the Sierra	INEC, 2010	0.2969
People born in the Oriente	INEC, 2010	0.259
Population density	INEC, 2010	0.2179
ED to deforested areas in 2002	Own data	0.4606
ED to roads	SIGTIERRAS	0.4606
ED to oil fields	PRAS 2016	0.2373

difference between the calculated accuracy using the validation data and expected accuracy if one were to randomly guess at the class memberships of the validation pixels. We obtained a training accuracy rate of nearby 80% and a skill measure of nearby 0.6. The only dynamic driver is ED to deforested areas and, still for simplification purposes, we have not used any constraint or incentive, although we have cadastral division and that our study area is crossed by the Yasuni National Park, in the east.

3 A Classic Markovian BAU Scenario and Its Adjustments

First, we performed a classic BAU trend scenario, as implemented by default in LCM, where future change quantities are estimated using Markov chains, based on two training dates. For this reason, we have named it “Markovian BAU”. In a second step, we proposed two consecutive adjustments to the Markovian BAU: an adjustment to exceed its linearity, taking into account a larger time period (BAU-a), and an adjustment to integrate demographic data (BAU-b).

These two adjustments are made by corrections to the Markov transition matrix (Table 4) and are intended to improve quantitative estimation of change intensity in trend scenarios as part of path-dependent pattern-based modelling approaches. In a third step, we considered the spatial aspects of these adjustments.

Table 6 Markovian BAU scenario simulated areas (ha)

Land cover	2017 by classification	2017 by Markovian BAU predicted area	Model deviations (%)
Water	0	5.67	
Forested areas	237400.83	231847.28	-2.3
Deforested areas	49218.48	54690.28	+11.1
Urban areas	123.21	110.03	-10.7

3.1 Markovian BAU

The classic Markovian BAU scenario uses the Markovian matrix calculated by LCM (Table 4) based on the training period we have defined (2002–2013) to determinate future quantities of change in simulations. Under this scenario, the hard simulation produced by LCM overestimates deforestation quantities: as shown in Table 6, nearly 54700 ha of deforestation are estimated by the simulation in 2017, compared to almost 49200 ha on the classification (Table 2), that is about 11.1% overestimation.

We assume that this overestimation is mainly due to the non-inclusion of the observed trend of increasing deceleration in the rate of deforestation, as LCM only estimated the quantities of changes from only two training dates. We try to correct this below, by modifying the matrix to take into account the deceleration trend.

3.2 BAU-a

The first adjustment we made is therefore to take into account a longer period of time for model calibration. We assume indeed that an observation of the dynamics prior to those of the strict training period (2002–2017) would allow us to better integrate the slowing of the rate of deforestation and thus limit the overestimation of deforestation quantities by the model, that we have previously observed. We have therefore, in concrete terms, changed the original transition matrix to better integrate this deceleration, by multiplying the transition probabilities that interested us by one factor: the ratio between the annual deforestation rates for the periods 1998–2002 and 2002–2013 (Table 3). This ratio, about 0.77, was therefore used to weight the transition probability from forested to deforested area, in bold in Table 7. We then adjusted the cell of persistence of the forested areas class accordingly, in such a way that the sum of the row equals 1 (difference between 1 and the new transition probability). This new modified matrix (Table 7) has been implemented in LCM and has led to new simulations, the results of which in terms of area by class are presented in Table 8.

Under this new scenario, we observe this time a lower overestimation of deforestation quantities by the model: LCM overestimates only 7.3%.

Table 7 Modified matrix of Markovian transition probabilities: the BAU-a scenario

	Water	Forested areas	Deforested areas	Urban areas
Water	0	0.3333	0.3333	0.3333
Forested areas	0	0.9736	0.0264	0
Deforested areas	0	0.0882	0.9106	0.0012
Urban areas	0	0	0.2143	0.7857

Table 8 BAU-a scenario simulated areas (ha)

Land cover	2017 by classification (ha)	2017 by BAU-a predicted area (ha)	Model deviations (%)
Water	0	5.67	
Forested areas	237400.83	233743.90	-1.5
Deforested areas	49218.48	52793.67	+7.3
Urban areas	123.21	110.03	-10.7

Table 9 Demographic data from the population census (INEC) and calculation of the ratio between annual growth rates

<i>Population of Orellana Canton</i>		
1990	2001	2010
19674	42010	72795
<i>Growth rates (%)</i>		
1990–2001	2001–2010	Ratio
10.32	8.14	0.79

3.3 BAU-b

Our second proposal to adjust the BAU trend scenario is to integrate population growth dynamics into the transition probability matrix, to make it more realistic. Indeed, population growth is often considered as a major driver of deforestation in the world, in Latin America and especially in NEA (Preston 1996; Armenteras et al. 2017; Jarrín-V. et al. 2017).

Therefore, using the available demographic data from the population censuses (INEC), we calculated a new ratio to reweight the transition matrix. On our study area, the only demographic data available at a fixed spatial scale over time, allowing the calculation of a population growth rate, were those at the cantonal level, and we focused on the canton of Orellana, which includes the *Dayuma parroquia* and most of our study area (Fig. 1). These data (Table 9) indicate that population growth slowed between 1990–2001 (10.32%) and 2001–2010 (8.14%), a trend effectively similar to that of the deforestation rate over a comparable period. We thus calculated the ratio between the annual population growth rates for the two periods (1990–2001 and 2001–2010). This ratio, of 0.79, was used to reweight the transition probability from forested to deforested area, this time in the BAU-a transition matrix (Table 7),

Table 10 Modified matrix of Markovian transition probabilities: the BAU-b scenario

	Water	Forested areas	Deforested areas	Urban areas
Water	0	0.3333	0.3333	0.3333
Forested areas	0	0.9792	0.0208	0
Deforested areas	0	0.0882	0.9106	0.0012
Urban areas	0	0	0.2143	0.7857

Table 11 BAU-b simulated areas (ha)

Land cover	2017 by classification (ha)	2017 by BAU-b predicted area (ha)	Model deviations (%)
Water	0	5.67	
Forested areas	237400.83	235088.43	-1
Deforested areas	49218.48	51449.13	+4.5
Urban areas	123.21	110.03	-10.7

in the same way as we did previously (i.e. by weighting the transition from forested to deforested areas by this new factor and then recalculating the other elements of the row). As before, the new matrix (Table 10), resulting from the calculation, was used in LCM to generate new simulations. The results in terms of quantities are presented in Table 11, in comparison with the surfaces of the classification.

As we can observe, after this second adjustment of the transition probability matrix, the overestimation by LCM is only 4.5% compared to the classification (Table 11). Our successive adjustments have therefore reduced the overestimation of quantities by more than half: whereas the classic Markovian BAU scenario simulated about 11.1% more deforestation than observed while the adjusted BAU-b scenario, the most advanced, generates only 4.5% of deviations to the model.

It seems that the adjustments have improved the quantitative estimation of change intensity, by exceeding the linearity of Markovian BAU scenarios based on only two dates and weighting the change probability matrix with demographic data. It is now a matter for us to briefly analyse the spatial effects of these adjustments.

3.4 Spatial Aspects

In order to consider the spatial aspects of our successive quantitative adjustments, that led to the simulation results of the BAU-b scenario presented before, we use here the method developed by Chen and Pontius (2010). Based on the observation that the Kappa indices are ineffective for accuracy assessment (Pontius and Millones 2011) on the one hand, and the need for statistical assessment on the other (Pontius et al. 2004), a part of this method consists in categorizing pixels into four categories in order to identify omission and commission errors (Pontius 2000): (i) correct due

Table 12 Overall prediction successes and error across the entire study area for Markovian BAU and BAU-b scenarios (%)

	NS	FA	H	DM	RM
Markovian BAU	89.10	1.32	0.89	4.38	4.31
BAU-b	89.40	1.03	0.71	4.56	4.31

NS null successes, *FA* false alarms, *H* hits, *DM* deforestation misses, *RM* reforestation misses

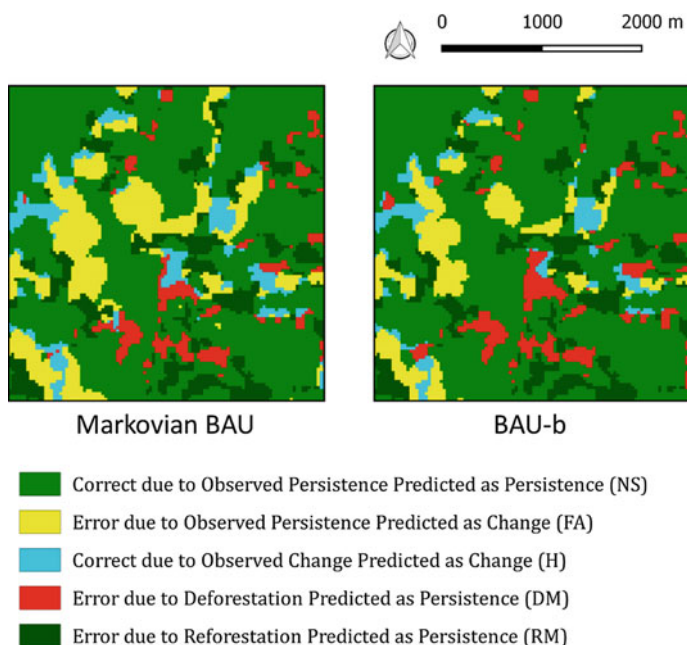


Fig. 2 Accuracy components based on observed land cover (2013, 2017) and 2017 predicted land cover maps from the Markovian BAU and the BAU-b

to observed persistence predicted as persistence (null successes), (ii) error due to observed persistence predicted as change (false alarms), (iii) correct due to observed change predicted as change (hits) and (iv) error due to observed change predicted as persistence (misses).

Table 12 shows the proportion of each of these categories in the Markovian BAU and in the BAU-b scenarios, calculated as a percentage of our study area. We have chosen to indicate separately errors due to reforestation (transition from deforested to forested area), which, as we recall, is not a process taken into account by the model. Figure 2 shows a portion of the territory simulated by the Markovian BAU scenario (left) and by BAU-b scenario (right), qualified according to this categorization of prediction successes and errors.

This analysis of the spatial aspects of simulation successes and errors demonstrate that quantitative adjustments to the probability matrices of change are not devoid of

spatial consequences. Indeed, as shown in Table 12, hits, which refer to deforestation that occurred between 2013 and 2017 and correctly predicted by LCM, represent nearby 0.9% of the study area before adjustments compared to almost 0.7% after adjustments (BAU-b). This is also visible on the map (Fig. 2) showing a representative detail of the study area, where the hits appear in blue: they are more numerous on the left extract assessing the Markovian BAU than on the right extract assessing the BAU-b. Inversely, errors due to deforestation predicted as persistence (DM) are increasing after adjustments: they represented nearly 4.4% of the study area before adjustments compared to nearly 4.6% after adjustments (in red on map extracts).

Then, if we relate the area of hits in both cases (Markovian BAU and BAU-b) to the deforestation area simulated by the two scenarios, we also find that hits are decreasing after successive adjustments. Under the Markovian BAU scenario, about 3200 ha of hits are observed among the 54690 ha of simulated deforestation, or approximately 5.9%. For BAU-b, about 2045 ha of hits are observed for 51450 ha of simulated deforestation, or almost 4%: the proportion of hits in simulated deforestation decreases as a result of adjustments.

4 Discussion

This paper shows that BAU scenarios as implemented in LCM, i.e. based on a training period established from only two dates, may be insufficient to provide a correct quantitative estimation of change intensity. Indeed, using a classic Markovian BAU scenario, the simple LCM model used here was not able to accurately reproduce the observed trend, i.e. a slowing down in the deforestation rate, since it overestimates quantities of deforestation. The work presented here explores two ways of improving change intensity prediction in BAU scenarios in land change modelling: extend the trend observation period and use thematic data to more accurately predict future quantities of change. These assumptions are applied by successive weightings of the model's transition probability matrix.

First of all, the results show that such adjustments of the probability transition matrix can improve path-dependant modelling approaches: they led to a lower over-estimation of deforestation quantities in the simulations. These results therefore highlight the value of incorporating a longer time period and the benefits of taking socio-demographic data into account during the calibration step, to exceed the linearity in the construction of change quantities prediction and to make them more realistic. This last approach is in line with the idea of "socializing" pixels, which appeared in the 1990s (Martin and Bracken 1993; National Research Council 1998).

These results also imply that BAU scenarios would benefit from being better designed: as these are so-called "trend" scenarios, because they are path-dependent approaches, we believe that it would be more efficient to build them on the basis of a more in-depth understanding of the trends, i.e. beyond the only two training dates allowed by the current software packages. In this sense, the integration of higher-order Markov chains (Ching et al. 2013) into LUCC modelling tools could

be a potential path to consider, because the successive adjustments of the Markovian matrices proposed here cannot constitute a robust methodology applicable to all cases and all types of thematic data.

However, these adjustments emphasize the need to integrate socio-economic data at each step of the LUCC modelling process, in one way or another and therefore not only at the spatial allocation of changes step as is currently the case in software packages. We believe that socio-economic data must be used to estimate future quantities of LUCC, as they are used to predict spatial allocation. This is obviously valid for all LUCC modelling approaches, above and beyond BAU scenarios. Because land use systems are characterized by multiple, non-linear and complex interactions between societies and environment, at different temporal and spatial scales (Geist et al. 2006), and cannot therefore be limited to the use of purely spatial or physical drivers, whatever the stage of the modelling process. In addition, it is likely that the coming decades will be characterized by the multiplication of accessible socio-economic data as well as those of big data. The latter represent a major challenge for many scientific disciplines and geography and geomatics are no exception (Kitchin 2013). Lastly, it is interesting to note that population projections studies have become increasingly numerous and accessible in recent years, including in Global South countries such as Ecuador, where they are produced and published by INEC. These data can be useful in the development of prospective BAU scenarios, especially when they are themselves trend-based.

Finally, the results show that in LCM, an improvement in the estimation of future quantities of change can lead to a decrease in the proportion of hits in predicted changes, i.e. changes that occurred and were correctly predicted. The improvement in the prediction of change quantities has indeed led to a decrease in the quality of the results in terms of spatial allocation in the case of the study presented here. This is simply because to spatially allocate changes, LCM selects the pixels with the highest change potential on the transition potential map calculated by the MLP neural network based on the relationship between the changes that occurred during the training period and drivers. But a reduction in quantities simply results in a smaller selection of pixels and therefore a smaller simulated change area. However, if the software simulates fewer changes, it is less likely to hit the target: the changes that occurred. Thus, this result suggests that quantitative improvements must be accompanied by progress in the spatial allocation of changes, in particular the reproduction of realistic patterns of change by models e.g. as Dinamica EGO software allows it better through its mechanism of expander/patcher (Soares-Filho et al. 2002; Rodrigues and Soares-Filho 2018). Process-based LUCC modelling approaches like agent-based models are also an interesting approach in this field (Parker et al. 2003; Matthews et al. 2007) and their coupling to pattern-based approaches is still an important scientific issue (Castella and Verburg 2007).

Although this approach allowed us to obtain results which contribute to a brief reflection on LUCC modelling practices and especially on the calibration stage of trend/BAU scenarios, it has however several limitations. These limitations are due both to the data used, to the choices made during the construction of the model in LCM and to the method by which we adjusted matrices.

Initially, these limitations concern the accuracy of the remotely sensed data as model input data. Indeed, it has been demonstrated that uncertainty is present at each step of the construction of land cover data and that is therefore significantly present in LUCC models (Garcia Alvarez 2018). Besides, this uncertainty is at the root of a numerous and ongoing work on improving satellite image classification techniques which show that there is always a scope for improvement (Lu and Weng 2007; Tso and Mather 2009). Nonetheless, it is important to remember that while supervised classification methods offer many advantages, including time savings, practical systematization and greater objectivity, their accuracy is often lower than manual classification by photo-interpretation. In brief, the results must be balanced according to the confidence we can place in land cover data, especially since they were not validated by field surveys.

Then, regarding the limitations inherent in the construction of the model, some choices made to simplify the model in order to improve understanding can be discussed. In particular, the non-inclusion of transitions from forested to deforested areas (reforestation process) can reduce model accuracy as much as the non-use of more dynamic data updated at each iteration like cadastral data or dynamic road modelling. However, the main purpose of this paper is to propose an improvement of the quantitative estimation of change intensity in trend scenarios, that is why the emphasis has been placed mainly on the quantitative aspects, to the detriment of certain details, which may nevertheless usually be essential for the development of a complete LUCC model.

Finally, the last limitation that we can highlight is the mismatch of spatial and temporal scales when the transition probability matrix has been modified. Indeed, for the second adjustment (the BAU-b scenario), we used cantonal demographic data for a model applied to a lower spatial level. Another bias lies in the fact that these cantonal data include several cities, characterized by specific demographic dynamics, while our territory is essentially rural. In addition, the time scale of the population censuses used to weight the matrix does not exactly match that of our classifications.

5 Conclusion

Based on a simple LUCC model developed with LCM, this work highlights the need to extend the trend observation period and to include thematic data in the calibration step of path-dependent pattern-based modelling approaches, to improve the quantitative estimation of change intensity. Indeed, the successive adjustments to the original Markov matrix of transition probabilities have minimized the model's overestimation of deforestation.

A quick spatial analysis of the results also recalls that improving the quantitative estimation of changes cannot be done independently of progress in the spatial allocation of changes.

References

- Armenteras D, Espelta JM, Rodríguez N, Retana J (2017) Deforestation dynamics and drivers in different forest types in Latin America: three decades of studies (1980–2010). *Glob Environ Change* 46:139–147. <https://doi.org/10.1016/j.gloenvcha.2017.09.002>
- Baynard CW, Ellis JM, Davis H (2013) Roads, petroleum and accessibility: the case of eastern Ecuador. *GeoJournal* 78:675–695. <https://doi.org/10.1007/s10708-012-9459-5>
- Bilborrow RE, Barbieri AF, Pan W (2004) Changes in population and land use over time in the Ecuadorian Amazon. *Acta Amaz* 34:635–647. <https://doi.org/10.1590/S0044-59672004000400015>
- Bromley R (1981) The colonization of humid tropical areas in Ecuador. *Singap J Trop Geogr* 2:15–26. <https://doi.org/10.1111/j.1467-9493.1981.tb00114.x>
- Brown LA, Sierra R, Southgate D, Labao L (1992) Complementary perspectives as a means of understanding regional change: frontier settlement in the Ecuador Amazon. *Environ Plan A* 24:939–961. <https://doi.org/10.1068/a240939>
- Camacho Olmedo MT, Mas JF (2018) Markov Chain. In: Camacho Olmedo MT, Paegelow M, Mas J-F, Escobar F (eds) *Geomatic approaches for modeling land change scenarios*. Springer International Publishing, Cham, pp 441–445
- Camacho Olmedo MT, Paegelow M, Mas JF, Escobar F (2018) Geomatic approaches for modeling land change scenarios. An introduction. In: Camacho Olmedo MT, Paegelow M, Mas J-F, Escobar F (eds) *Geomatic approaches for modeling land change scenarios*. Springer International Publishing, Cham, pp 1–8
- Carr DL (2003) Proximate population factors and deforestation in tropical agricultural frontiers. *Popul Environ* 25:585–612. <https://doi.org/10.1023/B:POEN.0000039066.05666.8d>
- Castella J-C, Verburg PH (2007) Combination of process-oriented and pattern-oriented models of land-use change in a mountain area of Vietnam. *Ecol Model* 202:410–420. <https://doi.org/10.1016/j.ecolmodel.2006.11.011>
- Chen H, Pontius RG (2010) Diagnostic tools to evaluate a spatial land change projection along a gradient of an explanatory variable. *Landsc Ecol* 25:1319–1331. <https://doi.org/10.1007/s10980-010-9519-5>
- Chhabra A, Geist H, Houghton RA, Haberl H, Braimoh AK, Vlek PLG, Patz J, Xu J, Ramankutty N, Coomes O, Lambin EF (2006) Multiple impacts of land-use/cover change. In: Lambin EF, Geist H (eds) *Land-use and land-cover change*. Springer, Berlin, Heidelberg, pp 71–116
- Ching W-K, Huang X, Ng MK, Siu T-K (2013) Higher-order Markov chains. *Markov chains*. Springer US, Boston, MA, pp 141–176
- Comber A, Balzter H, Cole B, Fisher P, Johnson S, Ogutu B (2016) Methods to quantify regional differences in land cover change. *Remote Sens* 8:176. <https://doi.org/10.3390/rs8030176>
- De Grande P (2016) El formato Redatam/The Redatam format. *Estudios Demográficos y Urbanos* 31:811. <https://doi.org/10.24201/edu.v31i3.15>
- Eastman J (2014) *TerrSet geospatial monitoring and modeling system*. Clark University, Worcester, MA
- Eastman JR, Toledano J (2018) A short presentation of the land change modeler (LCM). In: Camacho Olmedo MT, Paegelow M, Mas J-F, Escobar F (eds) *Geomatic approaches for modeling land change scenarios*. Springer International Publishing, Cham, pp 499–505
- Escobar F, van Delden H, Hewitt R (2018) LUCC scenarios. In: Camacho Olmedo MT, Paegelow M, Mas J-F, Escobar F (eds) *Geomatic approaches for modeling land change scenarios*. Springer International Publishing, Cham, pp 81–97
- Floriani LD, Magillo P (2009) Triangulated irregular network. In: Liu L, Özsu TT (eds) *Encyclopedia of database systems*. Springer US, Boston, MA, pp 3178–3179
- García Álvarez D (2018) *Aproximación al Estudio de la Incertidumbre en la Modelización del Cambio de Usos y Coberturas del Suelo (LUCC)*. Universidad de Granada

- Geist H, McConnell W, Lambin EF, Moran E, Alves D, Rudel T (2006) Causes and trajectories of land-use/cover change. In: Lambin EF, Geist H (eds) *Land-use and land-cover change*. Springer, Berlin, Heidelberg, pp 41–70
- Hiraoka M, Yamamoto S (1980) Agricultural development in the Upper Amazon of Ecuador. *Geogr Rev* 70:423. <https://doi.org/10.2307/214077>
- Houet T, Aguejidad R, Doukari O, Battaia G, Clarke K (2016) Description and validation of a “non path-dependent” model for projecting contrasting urban growth futures. *Cybergeo*. <https://doi.org/10.4000/cybergeo.27397>
- Houet T, Gourmelon F (2014) La géoprospective – Apport de la dimension spatiale aux démarches prospectives. *Cybergeo*. <https://doi.org/10.4000/cybergeo.26194>
- Houssou L (2016) Simulation sociale à base d’agents du comportement microéconomique des ménages en Amazonie équatorienne, face aux contaminations pétrolières, aux dynamiques économiques et aux politiques publiques. MSc thesis, Université nationale du Vietnam, Institut francophone international, Hanoi, Vietnam
- Jarrín-V. PS, Tapia Carrillo L, Zamora G (2017) Demografía y transformación territorial: medio siglo de cambio en la región amazónica de Ecuador/Demography and territorial transformation: half a century of change in the Amazonian Region of Ecuador. *Eutopía, Revista de Desarrollo Económico Territorial* 81. <https://doi.org/10.17141/eutopia.12.2017.2913>
- Juteau-Martineau G, Becerra S, Maurice L (2014) Ambiente, petróleo y vulnerabilidad política en el Oriente Ecuatoriano: ¿hacia nuevas formas de gobernanza energética? *América Latina Hoy* 67:119. <https://doi.org/10.14201/alh201467119137>
- Kitchin R (2013) Big data and human geography: opportunities, challenges and risks. *Dialogues Hum Geogr* 3:262–267. <https://doi.org/10.1177/2043820613513388>
- Lambin EF, Turner BL, Geist HJ, Agbola SB, Angelsen A, Bruce JW, Coomes OT, Dirzo R, Fischer G, Folke C, George PS, Homewood K, Imbernon J, Leemans R, Li X, Moran EF, Mortimore M, Ramakrishnan PS, Richards JF, Skånes H, Steffen W, Stone GD, Svedin U, Veldkamp TA, Vogel C, Xu J (2001) The causes of land-use and land-cover change: moving beyond the myths. *Glob Environ Change* 11:261–269. [https://doi.org/10.1016/S0959-3780\(01\)00007-3](https://doi.org/10.1016/S0959-3780(01)00007-3)
- Lu D, Weng Q (2007) A survey of image classification methods and techniques for improving classification performance. *Int J Remote Sens* 28:823–870. <https://doi.org/10.1080/01431160600746456>
- Maestriperi N, Paegelow M (2013) Validation spatiale de deux modèles de simulation: l’exemple des plantations industrielles au Chili. *Cybergeo*. <https://doi.org/10.4000/cybergeo.26042>
- Mahmood R, Pielke RA, Hubbard KG, Niyogi D, Dirmeyer PA, McAlpine C, Carleton AM, Hale R, Gameda S, Beltrán-Przekurat A, Baker B, McNider R, Legates DR, Shepherd M, Du J, Blanken PD, Frauenfeld OW, Nair US, Fall S (2014) Land cover changes and their biogeophysical effects on climate: land cover changes and their biogeophysical effects on climate. *Int J Climatol* 34:929–953. <https://doi.org/10.1002/joc.3736>
- Martin D, Bracken I (1993) The integration of socioeconomic and physical resource data for applied land management information systems. *Appl Geogr* 13:45–53. [https://doi.org/10.1016/0143-6228\(93\)90079-G](https://doi.org/10.1016/0143-6228(93)90079-G)
- Mas J (2004) Modelling deforestation using GIS and artificial neural networks. *Environ Model Softw* 19:461–471. [https://doi.org/10.1016/S1364-8152\(03\)00161-0](https://doi.org/10.1016/S1364-8152(03)00161-0)
- Mas J-F (1999) Monitoring land-cover changes: a comparison of change detection techniques. *Int J Remote Sens* 20:139–152. <https://doi.org/10.1080/014311699213659>
- Mas J-F, Kolb M, Paegelow M, Camacho Olmedo MT, Houet T (2014) Inductive pattern-based land use/cover change models: a comparison of four software packages. *Environ Model Softw* 51:94–111. <https://doi.org/10.1016/j.envsoft.2013.09.010>
- Mas JF, Paegelow M, Camacho Olmedo MT (2018) LUCC modeling approaches to calibration. In: Camacho Olmedo MT, Paegelow M, Mas J-F, Escobar F (eds) *Geomatic approaches for modeling land change scenarios*. Springer International Publishing, Cham, pp 11–25

- Mas J-F, Soares Filho B, Pontius R, Farfán Gutiérrez M, Rodrigues H (2013) A suite of tools for ROC analysis of spatial models. *ISPRS Int J Geo-Inf* 2:869–887. <https://doi.org/10.3390/ijgi2030869>
- Matthews RB, Gilbert NG, Roach A, Polhill JG, Gotts NM (2007) Agent-based land-use models: a review of applications. *Landsc Ecol* 22:1447–1459
- Mena CF, Bilsborrow RE, McClain ME (2006) Socioeconomic drivers of deforestation in the Northern Ecuadorian Amazon. *Environ Manage* 37:802–815. <https://doi.org/10.1007/s00267-003-0230-z>
- Morin L (2015) Diagnostic agraire d'un front pionnier en Amazonie équatorienne, Paroisse de Dayuma, province d'Orellana, Equateur. MSc thesis, SupAgro Montpellier IRC
- National Research Council (1998) People and pixels: linking remote sensing and social science. National Academies Press, Washington, D.C.
- Oliver TH, Morecroft MD (2014) Interactions between climate change and land use change on biodiversity: attribution problems, risks, and opportunities: interactions between climate change and land use change. *Wiley Interdiscip Rev Clim Change* 5:317–335. <https://doi.org/10.1002/wcc.271>
- Orme CDL, Davies RG, Burgess M, Eigenbrod F, Pickup N, Olson VA, Webster AJ, Ding T-S, Rasmussen PC, Ridgely RS, Stattersfield AJ, Bennett PM, Blackburn TM, Gaston KJ, Owens IPF (2005) Global hotspots of species richness are not congruent with endemism or threat. *Nature* 436:1016–1019. <https://doi.org/10.1038/nature03850>
- Paegelow M (2018) Impact and integration of multiple training dates for Markov based land change modeling. In: Camacho Olmedo MT, Paegelow M, Mas J-F, Escobar F (eds) *Geomatic approaches for modeling land change scenarios*. Springer International Publishing, Cham, pp 121–138
- Paegelow M, Camacho Olmedo MT (2010) Modelos de simulación espacio-temporal y Teledetección: el método de la segmentación para la cartografía cronológica de usos del suelo. *Serie Geográfica – Universidad de Alcalá*, pp 19–34
- Paegelow M, Camacho Olmedo MT, Mas J-F, Houet T (2014) Benchmarking of LUCC modelling tools by various validation techniques and error analysis. *Cybergeogeo*. <https://doi.org/10.4000/cybergeogeo.26610>
- Parker DC, Manson SM, Janssen MA, Hoffmann MJ, Deadman P (2003) Multi-agent systems for the simulation of land-use and land-cover change: a review. *Ann Assoc Am Geogr* 93:314–337. <https://doi.org/10.1111/1467-8306.9302004>
- Perz SG, Walker RT (2002) Household life cycles and secondary forest cover among small farm colonists in the Amazon. *World Dev* 30:1009–1027
- Pontius GR (2000) Quantification error versus location error in comparison of categorical maps. *Photogramm Eng Remote Sens* 66:1011–1016
- Pontius RG, Huffaker D, Denman K (2004) Useful techniques of validation for spatially explicit land-change models. *Ecol Model* 179:445–461. <https://doi.org/10.1016/j.ecolmodel.2004.05.010>
- Pontius RG, Millones M (2011) Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *Int J Remote Sens* 32:4407–4429. <https://doi.org/10.1080/01431161.2011.552923>
- Pontius RG, Schneider LC (2001) Land-cover change model validation by an ROC method for the Ipswich watershed, Massachusetts, USA. *Agric Ecosyst Environ* 85:239–248. [https://doi.org/10.1016/S0167-8809\(01\)00187-6](https://doi.org/10.1016/S0167-8809(01)00187-6)
- Preston SH (1996) The effect of population growth on environmental quality. *Popul Res Policy Rev* 15:95–108. <https://doi.org/10.1007/BF00126129>
- Rodrigues H, Soares-Filho B (2018) A short presentation of Dinamica EGO. In: Camacho Olmedo MT, Paegelow M, Mas J-F, Escobar F (eds) *Geomatic approaches for modeling land change scenarios*. Springer International Publishing, Cham, pp 493–498
- Soares-Filho BS, Coutinho Cerqueira G, Lopes Pennachin C (2002) Dinamica—a stochastic cellular automata model designed to simulate the landscape dynamics in an Amazonian colonization frontier. *Ecol Model* 154:217–235. [https://doi.org/10.1016/S0304-3800\(02\)00059-5](https://doi.org/10.1016/S0304-3800(02)00059-5)
- Tso B, Mather P (2009) *Classification methods for remotely sensed data*, 2nd edn. CRC Press

- Veldkamp A, Lambin E (2001) Predicting land-use change. *Agric Ecosyst Environ* 85:1–6. [https://doi.org/10.1016/S0167-8809\(01\)00199-2](https://doi.org/10.1016/S0167-8809(01)00199-2)
- Wasserstrom R, Southgate D (2013) Deforestation, Agrarian reform and oil development in Ecuador, 1964–1994. *Nat Resour* 04:31–44. <https://doi.org/10.4236/nr.2013.41004>

Part IV
Spatial Scale as a Common Thread
in Geoinformation Analysis
and Modeling

Market Area Delineation for Airports to Predict the Spread of Infectious Disease



Carmen Huber and Claus Rinner

Abstract Air travel facilitates the international spread of infectious disease. While global air travel data represent the volume of travel between airports, identifying which airport an infected individual might use, or where a disease might spread after an infected passenger deplanes, remains a largely unexplored area of research and public health practice. This gap can be addressed by estimating airport catchment areas. This research aims to determine how existing market area delineation techniques estimate airport catchments differently, and which techniques are best suited to anticipate where infectious diseases may spread. Multiple techniques were tested for airports in the Province of Ontario, Canada: circular buffers, drive-time buffers, Thiessen polygons, and the Huff model, with multiple variations tested for some techniques. The results were compared qualitatively and quantitatively based on spatial patterns as well as area and population of each catchment area. There were notable differences, specifically between deterministic and probabilistic approaches. Deterministic techniques may only be suitable if all airports in a study area are similar in terms of attractiveness. The probabilistic Huff model appeared to produce more realistic results because it accounted for variation in airport attractiveness. Additionally, the Huff model requires few inputs and therefore would be efficient to execute in situations where time, resources, and data are limited.

Keywords Airport catchments · Huff model · Infectious disease · Public health · Retail geography

C. Huber · C. Rinner (✉)
Department of Geography and Environmental Studies, Ryerson University, 350 Victoria Street,
Toronto, ON M5B 2K3, Canada
e-mail: crinner@ryerson.ca

© Springer Nature Switzerland AG 2020
P. Kyriakidis et al. (eds.), *Geospatial Technologies for Local and Regional Development*, Lecture Notes in Geoinformation and Cartography,
https://doi.org/10.1007/978-3-030-14745-7_15

1 Research Background

Air travel is a key mechanism facilitating the international spread of infectious disease. The increasing volume and ease of air travel promotes the dispersal of pathogens and importation of infectious diseases worldwide. This poses a significant threat to environmental conservation and public health (Golnar et al. 2016; Hatcher et al. 2012; Kilpatrick et al. 2006; Tatem et al. 2006). Severe acute respiratory syndrome (SARS), West Nile virus and Zika virus are well-known examples of this phenomenon (Bird and McElroy 2016; Bogoch et al. 2016a, b; Fauci and Morens 2016; Golnar et al. 2016; Lounibos 2002; Powers 2015; The SARS Commission 2006). The majority of processes that aim to address infectious disease threats are reactive in nature (Kilpatrick et al. 2006). In response to events such as the outbreaks of SARS, West Nile and Zika, researchers have increasingly studied the global movement of humans to *predict* and possibly *prevent* the emergence of infectious disease. However, this has proven to be notoriously difficult, and many approaches involve numerous assumptions and analyses based on incomplete data (Kilpatrick et al. 2006). Despite these inherent challenges, researchers continue to find new ways to provide necessary support to decision-makers (Golnar et al. 2016).

Multiple studies have identified human air travel as one of the most important pathways for the importation of infectious disease to new areas (Golnar et al. 2016; Kilpatrick et al. 2006). Data on passenger and flight volumes have been used to model international connections and anticipate where infectious diseases might be imported. For example, Bogoch et al. (2016a) analyzed air travel volumes for travelers departing Brazil, to anticipate the international spread of Zika virus during the 2016 outbreak. Such analyses aim to help health care workers anticipate whether they may see travel-related infectious disease cases, and support public health officials in guiding resource distribution (e.g. for screening or communication campaigns) to locations at risk for local transmission if a case was imported. In fact, Zika virus was presumably imported to Miami from Brazil via air travel, and local transmission was initiated due to the presence of *Aedes* mosquitos and suitable environmental conditions (Bogoch et al. 2016a, b; Centers for Disease Control and Prevention, 2017).

The processes of infectious disease importation and spread are complex, but simplified modelling efforts can be effective (Golnar et al. 2016). While analysis using global air travel data indicates volume of travel between airports, identifying which airport an infected individual might use, or where a disease might spread after an infected passenger deplanes, remains a largely unexplored area of research. This gap can be addressed by estimating the area from which an airport attracts its passengers, referred to as its catchment area (Lin et al. 2016). In the absence of observed data, airport catchment areas can be estimated through a variety of models. While many catchment area techniques were developed for trade area analysis in retail geography (Boots 1980; Huff 2003; Huff and Black 1997; Reilly 1931), they have also been applied to model service areas for transportation (Debrezion et al. 2009; Lin et al. 2016; Wittman 2014). Existing methods can be classified into three main cate-

gories: deterministic, probabilistic, and customer profiling (Hernandez et al. 2004). Only deterministic and probabilistic methods are relevant to this study, since data representing spatial concentrations of airport users, which would be necessary for customer profiling, are lacking.

Deterministic approaches make a clear-cut assumption about the spatial dimension of the trade area (Hernandez et al. 2004). Trade areas are polygons that have definite boundaries, and assume that all customers come from the defined catchment (Hernandez et al. 2004). Proximity-only models are included in the deterministic category and include circular buffers or drive-time buffers along a road network (Hernandez et al. 2004; Lin et al. 2016). Another deterministic technique calculates Thiessen polygons (also known as Voronoi polygons) around sites (Boots and South 1997; Hernandez et al. 2004). Here, every customer is assigned to the closest site based on Euclidean distance. Customers are assigned to only one site, and the mid-points between sites form the trade area boundaries. The Thiessen polygon method has been adapted to incorporate weights based on store/site attractiveness (Boots 1980; Hernandez et al. 2004). While deterministic methods such as circular buffers have been used to estimate airport catchment areas for the purpose of anticipating spread of infectious disease (Brent et al. 2018), they may over-simplify the problem (Cervero et al. 1995; Debrezion et al. 2009; Lin et al. 2016; Sanko and Shoji 2009).

In contrast to deterministic approaches, probabilistic approaches do not assume that customers always choose the closest option, and therefore assign customer groups (households, census tracts, neighbourhoods) partially to the alternative sites (Hernandez et al. 2004). A widely-used probabilistic model is the Huff model (Huff 2003), which defines catchment or trade areas as series of zonal probability contours (Huff 1963). The Huff model is popular in retail geography because it is relatively straightforward to apply, conceptually appealing, and applicable to a wide range of problems (Huff 2003; Huff and Black 1997).

The Huff model results represent the probability of the population at each origin location to patronize each alternative service location. The two basic parameters of the model are attraction and distance, both of which have explicit behavioural bases (Huff and Black 1997). The attraction parameter represents the “impact of store size on consumer patronage for a given product when distance is held constant” (Huff and Black 1997). For certain products or services, size (or other associated attractiveness measure) is more important to consumers, and therefore would more greatly impact their choice alternative. The distance decay parameter represents the consumer’s willingness to travel for different types of products (Huff 1963). The choice set is another critical element of the Huff model. In a choice situation, there exists a universal set of alternative sites from which a consumer selects a subset based on their individual preferences. For example, some choice alternatives may be beyond a maximum distance the consumer is willing to travel (Huff and Black 1997). Specifying an accurate choice set is essential to minimizing prediction errors (Huff and Black 1997).

When selecting a method to estimate airport catchment areas, it is important to consider existing knowledge of how airports are used. For example, Debrezion et al. (2009) found that less than half of passengers at a Dutch railway survey chose their

nearest train station (Lin et al. 2016). Leon (2011) found that airline travellers in North Dakota will not use the local airport but instead use the competing major hub airport located 250 miles away. There is general consensus that consumers are willing to travel further to reach a more desirable location, but it is difficult to determine the maximum distance they would be willing to travel especially considering this distance likely changes between regions (Leon 2011; Lin et al. 2016). These examples suggest that for the application of delineating catchment areas for transportation, deterministic proximity-only models may be too coarse (Debrezion et al. 2009; Leon 2011; Lin et al. 2016; Wittman 2014), although circular buffers have been used frequently to define airport catchment areas (Bilotkach et al. 2012; McLay and Reynolds-Feighan 2006; Wang 2000; Wittman 2014). Lin et al. (2016) suggest that gravity models (such as the Huff model) may be a more appropriate approach than proximity-only approaches, since they incorporate not only distance but also attraction. The main considerations of the Huff model align with what studies have shown to be the greatest determinants of airport choice (Başar and Bhat 2004; Hess and Polak 2005; Ishii et al. 2009; Leon 2011; Suzuki 2007).

While some localized analyses have been conducted to model airport catchment areas (Augustyniak and Olipra 2014; Lieshout 2012), their techniques would be complex to apply at a national, or global scale. For example, if a drive-time distance or cut-off were incorporated in a model, it might be unrealistic to apply a single appropriate distance to the entire study area, where characteristics of the population and environment likely differ widely (Lin et al. 2016; Upchurch et al. 2004).

To support rapid response to infectious disease outbreaks, we explore the differences in how the multiple available methods estimate airport catchment areas. This research aims to answer two questions: How do various market area delineation techniques estimate airport catchments differently? And, which techniques are best suited to anticipate where infectious diseases may spread internationally? Airports in the Province of Ontario, Canada, served as a test case for which to compare results between techniques.

2 Data and Methods

The case study of Ontario, Canada, was selected because of the province's large territory and high population, and the relevant context of the outbreak of severe acute respiratory syndrome (SARS) in Toronto, Ontario, in 2003. Ontario is the second-largest province in Canada and home to over 13.5 million people (Government of Ontario 2018). The highest population densities are clustered around Toronto with smaller clusters of moderately-high population density near London, Hamilton, and Ottawa (Fig. 1). Ontario's 77 airports are also concentrated in the south. Figure 1 shows the eight major airports that had 2016 passenger volumes reported by Statistics Canada.

Ontario's high proportion of foreign-born population (29%) also makes the province relevant to this study (Ontario Ministry of Finance 2017). This high pro-



Fig. 1 Population density and locations of Ontario’s major airports by 2016 passenger volume (Statistics Canada 2016a)

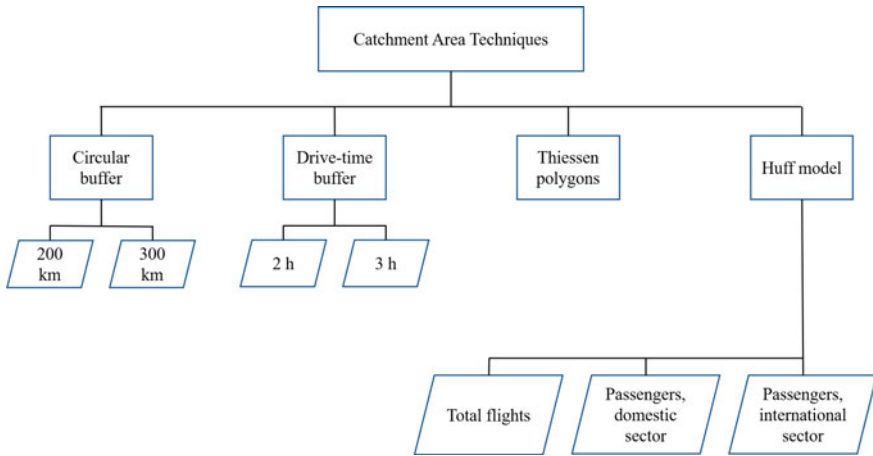


Fig. 2 Catchment area delineation techniques applied to Ontario airports, with input specifications for circular buffer radius, drive-time extent, and the Huff model’s airport attractiveness metric

portion is indicative of a high volume of transnational migrants, and increased frequency of cross-border travel for personal, education, and business purposes (Levitt and Jaworsky 2007). This phenomenon adds to the magnitude of international air travel to and from Ontario. The concern of importation of infectious disease via air travel is especially relevant to Toronto, in which immigrants make up close to one half of the city’s population. In 2003, Toronto experienced the largest outbreak of SARS outside of Asia (Bell 2004; Muller et al. 2006; Summers 2013). Bell (2004) noted “limiting the spread of infection by domestic and international travel” as one of the interventions that aided in containing this outbreak in some parts of the world. The importation of SARS to Toronto via international travel is an example of where interventions failed and had a devastating impact on both human health and the economy (The SARS Commission 2006).

The catchment area delineation techniques applied to Ontario’s airports included three deterministic techniques (circular buffers, drive-time buffers, and Thiessen polygons) and one probabilistic technique (Huff model) (Fig. 2), with some variations based on parameter inputs. Each technique had unique data requirements (Fig. 3). Results from each technique were compared and evaluated based on the applicability to analyzing the potential for international spread of infectious disease. We assume that these methods are more representative of Ontario residents than tourists or visitors, who might take longer routes to reach various tourist attractions.

Data on air passenger traffic and flights for Ontario airports were obtained from Statistics Canada (2016a). Only eight of 77 airports had associated passenger traffic and flights data, as detailed in Table 1.

To estimate catchment areas using circular buffers the only necessary input was the spatial locations of the airports. Buffers were produced based on two distances: 200 and 300 km. Brent et al. (2018) applied a 200 km buffer to airports in their

Table 1 Airport passenger and flight volumes for 2016, where ‘-1’ indicates data were unavailable or suppressed (Statistics Canada 2016a)

Name	Code	Pass. enplaned	Pass. deplaned	Total pass.	Domestic sector	Transborder sector	Intl. sector	Pass. flights
Toronto/Lester B Pearson	YYZ	21,347,368	21,451,589	42,798,957	16,741,872	11,572,354	14,484,731	410,316
Ottawa/Macdonald-Cartier Intl.	YOW	2,303,278	2,304,815	4,608,093	3,607,040	611,089	389,964	72,787
Thunder Bay	YQT	377,176	376,445	753,621	-1	-1	-1	27,463
London	YXU	257,516	250,689	508,205	467,075	17,758	23,372	10,277
Timmins	YTS	104,299	103,555	207,854	207,854	0	0	9475
Sudbury	YSB	113,888	113,315	227,203	-1	-1	0	8952
Windsor	YQG	160,068	159,254	319,322	-1	-1	-1	6970
Hamilton	YHM	160,255	159,009	319,264	263,802	-1	-1	3816

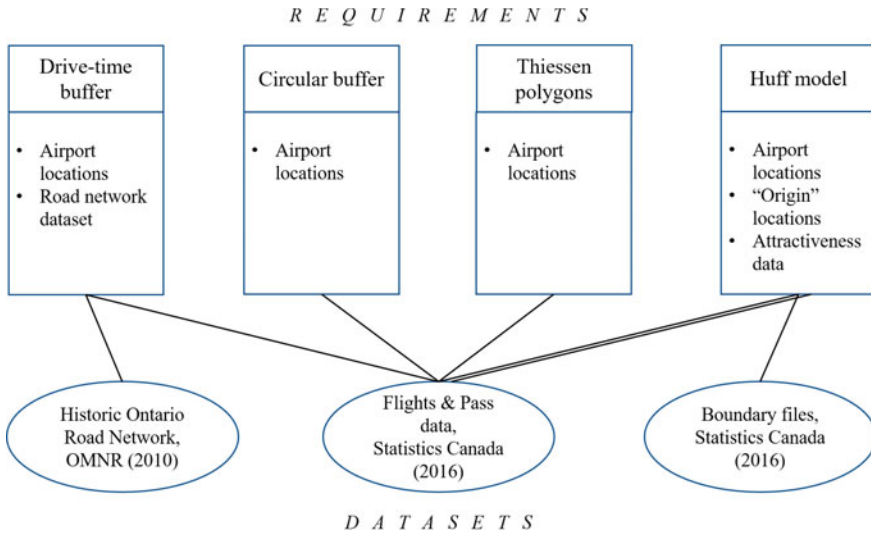


Fig. 3 Data requirements and sources by catchment area delineation technique (double line indicates data source was used to fulfill two data requirements)

analysis of potential international spread of Yellow Fever globally. They used flight itinerary data of travelers who departed yellow fever-endemic areas. 200 km served as the lowest buffer distance. While visualized as full circles on the resulting map, the area and population associated with each buffer was calculated based only on the portion that intersected the Ontario boundary (including smaller inland water bodies but excluding the Great Lakes and Hudson’s Bay).

The drive-time buffers approach required an additional dataset representing the road network. A road network dataset was created for Ontario using a road network file from the Ontario Ministry of Natural Resources (2010). Since this road network covered Ontario only, resulting buffers were automatically restricted to provincial land. Buffers were produced based on two drive-time cut-offs: two and three hours. A maximum driving-distance to reach an airport in Ontario could not be referenced from existing literature. Thus, a 2-h cut-off was selected because preliminary exploration revealed that it was generally comparable in extent to the 200 km buffer for YYZ—the airport with the highest passenger volume (Table 1). A 3-h cut-off was also applied to include a larger drive-time cut-off for comparison.

Like the circular buffer technique, Thiessen polygons required the single input of airport locations. Thiessen polygons form a tessellation that exhaustively fills the study area and do not overlap. Locations that are equally close to more than one airport collectively form the boundaries of the Thiessen polygons (Yamada 2016). In this study, each Thiessen polygon represented the catchment area of the respective airport.

The Huff model was the only probabilistic technique that was tested. It estimates catchment areas using both distance to and attractiveness of each airport, while also

incorporating distance to and attractiveness of all other airports. The Huff model is described by:

$$P_{ij} = \frac{S_j^\alpha / D_{ij}^\beta}{\sum_{i=1}^n S_j^\alpha / D_{ij}^\beta}$$

where P_{ij} is the probability of an individual located at i choosing airport j , S_j is a measure of attractiveness for j , D_{ij} is the distance from i to j , α is the attractiveness exponent and β is the distance decay exponent.

Probabilities were calculated for census subdivisions, which correspond to the municipalities (Statistics Canada 2016b), the level at which many public health programs and procedures are operationalized. To represent airport attractiveness, we tested multiple variables including total flights, total domestic passengers, and total international passengers. These data were obtained from the same Statistics Canada dataset that included airport locations. Parameterization of the Huff model was based on findings from a related study by one of the authors. On this basis, a value of 2 was applied as the distance decay exponent (beta), while no attractiveness exponent was applied (i.e., alpha = 1).

For all spatial analysis, modeling, and mapping, an open-source software package, QGIS, was used in conjunction with the “Location Analytics” toolset, which is under development (https://github.com/ryersongeo/qgis_location_analytics).

The results were visually compared within and between catchment area delineation techniques. Results were also quantitatively compared based on total area and population within each catchment area. For the Huff model, a probability threshold had to be defined to indicate which census subdivisions should be included in the area calculation. A minimum probability of 20% was selected to define the boundary of each catchment area. This threshold is similar to thresholds used to define market areas in the retail sector (Dolega et al. 2016). Population within each catchment area was calculated using population totals by census subdivision for 2016, obtained from Statistics Canada (2016c). For circular buffers, drive-time buffers, and Thiessen polygons, the total population within the catchment area was summed. If a subdivision was split by the catchment area boundary, the population of that subdivision was split proportionally based on area. For Huff model results, the population of each subdivision in the catchment area was multiplied by the probability of using each airport. The relationship between area and population for estimated catchment areas based on each technique was analyzed as an indicator of the spatial patterns of the risk of disease spread.

3 Analysis and Results

As expected based on the inherent characteristics of each technique, the estimated catchment areas notably differed. Results of each technique are shown in Figs. 4,

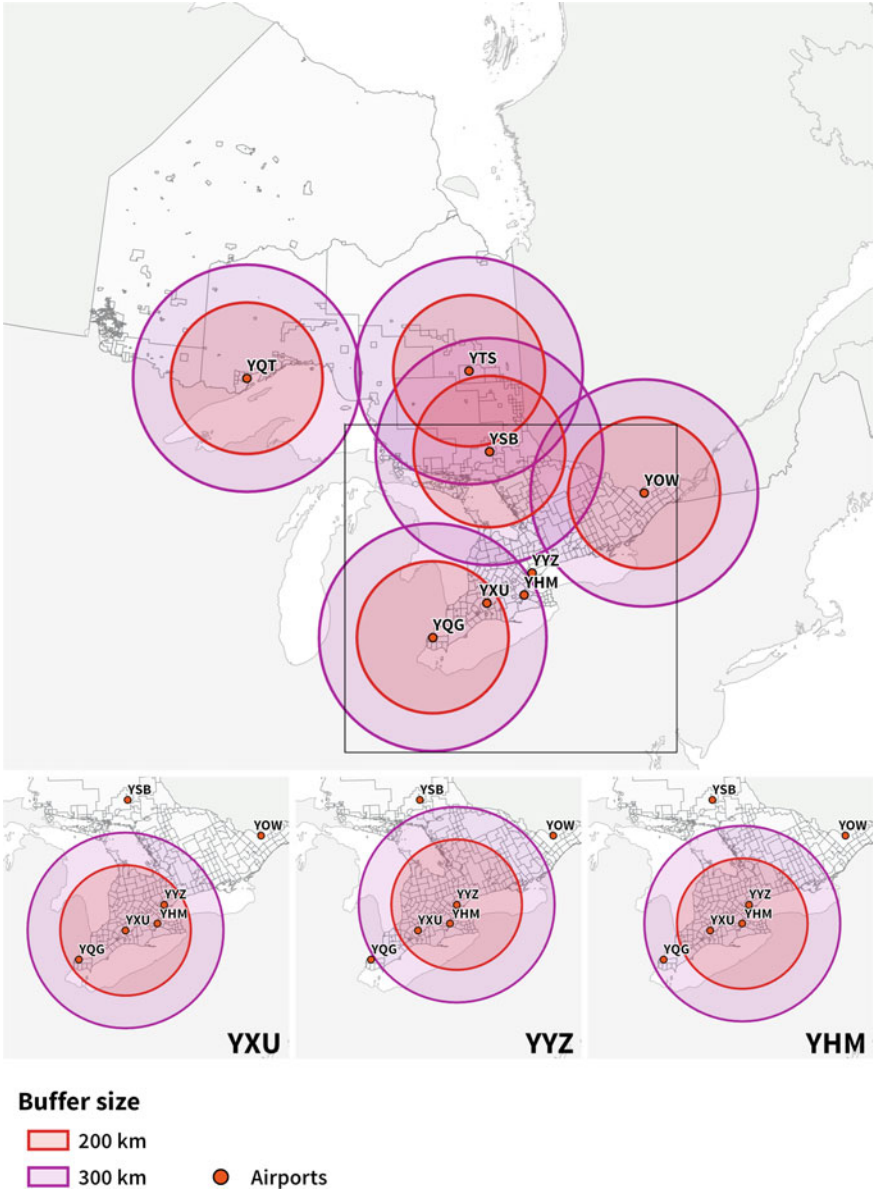


Fig. 4 Circular buffers for major Ontario airports overlaid on census subdivision boundaries. Insets provided for the London, Toronto, and Hamilton airports in Southern Ontario

6, 8, 10, 11 and 12, which reveal qualitative differences in the shape and extent of catchment areas. Quantitative differences in the estimated area and population associated with each catchment are shown in Table 2.

Table 2 Area (km²) and population for airport catchment areas as defined by different market area delineation methods

Airport code	Circular buffers					
	200 km		300 km			
	Area	Population	Area	Population		
YYZ	68,716	10,414,146	118,971	11,054,816		
YOW	40,899	1,752,100	69,416	2,441,445		
YQT	61,441	135,615	139,409	166,984		
YXU	48,540	9,945,855	70,582	10,885,048		
YQG	16,654	1,298,391	38,038	3,998,808		
YHM	56,291	10,295,559	93,135	11,183,121		
YSB	92,110	414,373	180,120	1,467,060		
YTS	112,561	125,490	218,602	397,199		
Airport code	Drive-time buffers				Thiessen polygons	
	2 h		3 h			
	Area	Population	Area	Population	Area	Population
YYZ	39,248	9,663,296	72,801	10,573,212	40,378	7,366,464
YOW	20,304	1,415,375	34,790	1,669,496	40,487	1,707,563
YQT	10,962	112,798	21,605	116,650	497,319	227,353
YXU	33,978	5,818,639	49,786	10,210,489	19,072	982,067
YQG	9116	905,679	20,492	1,999,344	5161	549,935
YHM	32,642	9,309,201	59,500	10,385,373	8072	2,050,521
YSB	7805	118,815	24,820	234,147	80,232	452,685
YTS	12,394	39,614	29,551	82,002	293,508	111,906
Airport code	Huff model					
	Flights		Passengers, domestic sector		Passengers, intl. sector	
	Area	Population	Area	Population	Area	Population
YYZ	961,118	10,285,369	968,352	10,812,078	981,143	12,146,910
YOW	35,000	1,366,554	44,206	1,397,449	17,137	1,032,229
YQT	143,910	162,333	N/A	N/A	N/A	N/A
YXU	6457	394,297	7638	410,131	867	91,049
YQG	2046	349,831	N/A	N/A	N/A	N/A
YHM	1152	123,331	1183	180,212	N/A	N/A
YSB	12,628	146,714	N/A	N/A	0	0
YTS	9642	55,226	8945	49,541	0	0

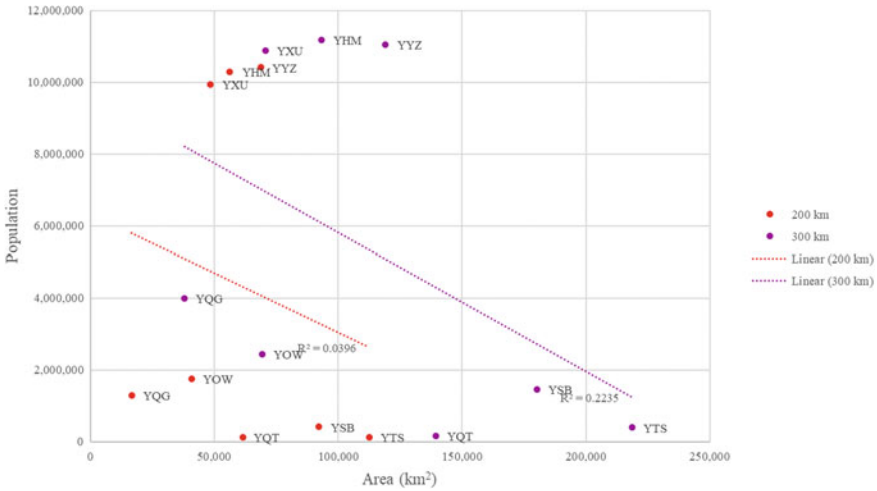


Fig. 5 Area and population for estimated airport catchment areas based on 200 and 300 km circular buffers

3.1 Circular Buffers

Despite the drastic difference in the number of annual passengers and flights between airports, the catchment sizes based on the circular buffers approach are equal by definition. The full circular buffers (shown in Fig. 4) were 125,535 km² for the 200 km buffer, and 282,453 km² for the 300 km buffer. Catchment areas were not impacted by the presence of other proximate airports—though this could have been incorporated if buffers were truncated.

The area and population for the portions of each catchment that were within the study area are detailed in Table 2. YTS was the largest in area for both the 200 km and 300 km buffer sizes at 112,561 km² and 218,602 km², respectively. YYZ and YHM had the largest population sizes based on both the 200 and 300 km circular buffers.

Figure 5 reveals that based on circular buffers, a larger catchment in terms of area was not necessarily associated with a larger population. While area was impacted by the proportion of the circular buffer that fell within the land boundaries of Ontario, the spatial distribution of population in the province meant that in many cases, circular buffers that had relatively small proportions of area falling within Ontario had the highest population sizes, and vice versa. For example, YYZ and YHM fell within the middle of all airports based on area but had the largest population. In contrast, YTS had the largest area but had one of the smallest populations. While the trend line in Fig. 5 shows that area was negatively associated with population, the relationship was weak and not statistically significant ($r = -0.199, p = 0.637$ for 200 km, and $r = -0.437, p = 0.237$ for 300 km).

3.2 Drive-Time Buffers

Generally, the size of catchment areas based on 2-h and 3-h drive-time buffers was similar between airports (Fig. 6), but could be categorized into two main groups based on whether they were located in areas with relatively high or low population density. Figure 1 shows that YQT, YTS, and YSB are surrounded by census subdivisions with relatively low population density. The road network is also less dense here. For these airports, the extent of the catchment area formed a web-like shape around the major roads. Thus, the catchment areas covered less area. These airport catchments were less than 13,000 km² based on a 2-h drive-time, and less than 30,000 km² based on a 3-h drive-time (Table 2).

In contrast, Fig. 1 shows that YYZ, YOW, YXU, YQG, and YHM are surrounded by census subdivisions with relatively high population density. The road network is also denser here so the catchment areas formed a fuller polygon. The area of the catchments for four of these five airports (excluding YQG) was between 20,000 and 40,000 km² based on a 2-h drive-time and between 34,000 and 73,000 km² based on a 3-h drive-time (Table 2). For airports surrounded by higher population density, catchments based on drive-time were similar to the circular buffers in extent.

As shown in Fig. 7 and in contrast to circular buffers, based on the drive-time buffers approach there was a strong, positive correlation between area and population that was significant based on Pearson's correlation coefficients ($r = 0.942$, $p = 0.000$ for 2-h, and $r = 0.919$, $p = 0.001$ for 3-h). For both buffer sizes, the airports with the largest catchments in terms of area also had the largest populations. As with circular buffers, YYZ and YHM had the largest catchment areas and associated populations.

3.3 Thiessen Polygons

In contrast to the two buffer approaches, there is great variation in the area of catchments based on Thiessen polygons (Fig. 8). Since there were more airports located in southern Ontario than in the rest of the province, catchment areas in southern Ontario were much smaller. Airports in less densely populated parts of Ontario, where fewer airport options existed, had much larger catchment areas. For example, Table 2 indicates that YQT located in northern Ontario had a catchment area of 497,319 km² while YYZ located near Toronto only had an area of 40,378 km². Both YQT and YTS had large catchment areas that extended to the northern boundary of the province.

Variation in population seemed to follow the opposite trend as variation in area. The airports in southern Ontario (plus YOW) tended to have larger population sizes associated with them even though the catchment areas were smaller. For example, Table 2 shows that a population of 227,353 fell within the catchment of YQT (northern Ontario), while a population of 7,366,464 fell within that of YYZ (southern Ontario). This follows the spatial pattern of population density in the province.

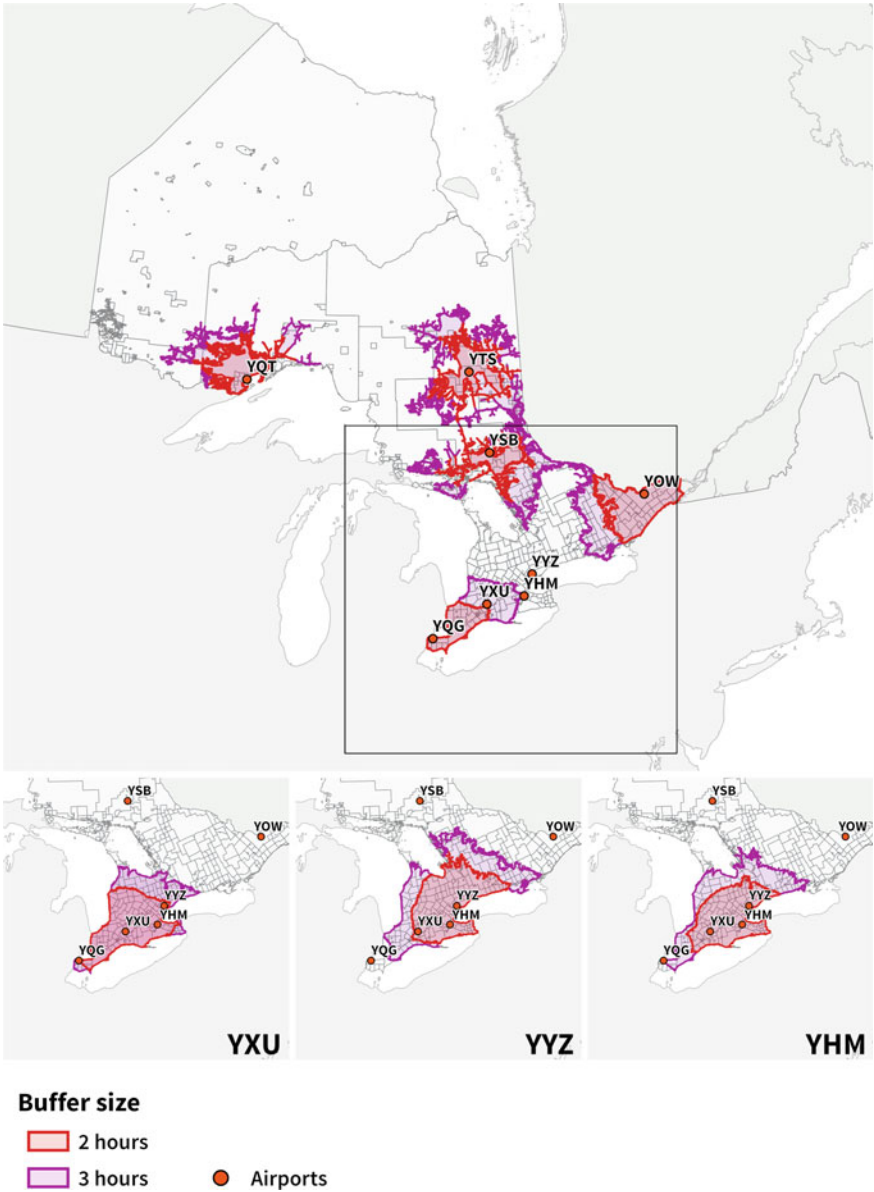


Fig. 6 Drive-time buffers for major Ontario airports overlaid on census subdivision boundaries

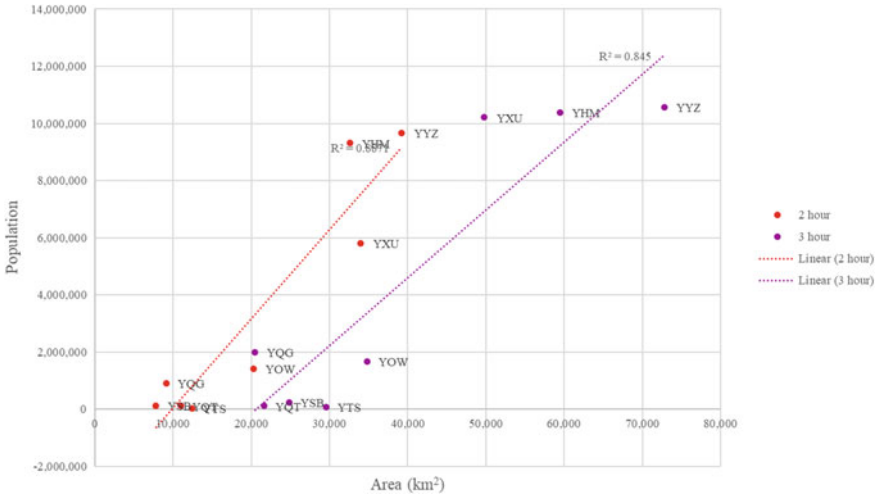


Fig. 7 Area and population for estimated airport catchment areas based on 2-h and 3-h drive-time buffers

These findings are confirmed by Fig. 9, which shows that the airports with the larger catchments in terms of area did not necessarily have large populations. There was a negative correlation between area and population, though it was weak and not statistically significant ($r = -0.356, p = 0.386$). YYZ is most notable in Fig. 9, because it had a relatively small area but has a population much higher than any other airport.

3.4 Huff Model

Estimated catchment areas based on the Huff model are shown in Figs. 10, 11 and 12, with airport attractiveness represented by total passenger flights, total domestic passengers, and total international passengers, respectively. Across all three variations of the Huff model, the most obvious pattern is that most census subdivisions were at least 40% likely to use YYZ over all other airports. Another notable characteristic of the Huff model results was that subdivisions located near one airport were less likely to use any other airport. While circular and drive-time buffers did not account for proximity to other airport options at all, Thiessen polygons arguably over-accounted for proximity to other airports by defining catchments based on only the midpoint between airport locations.

Between the Huff models based on total passenger flights, total domestic passengers, and total international passengers, very similar spatial patterns resulted. In all three cases, YYZ had the largest catchment, YOW had a moderately-large catchment, and all other airports were small or non-existent. A notable difference in results based

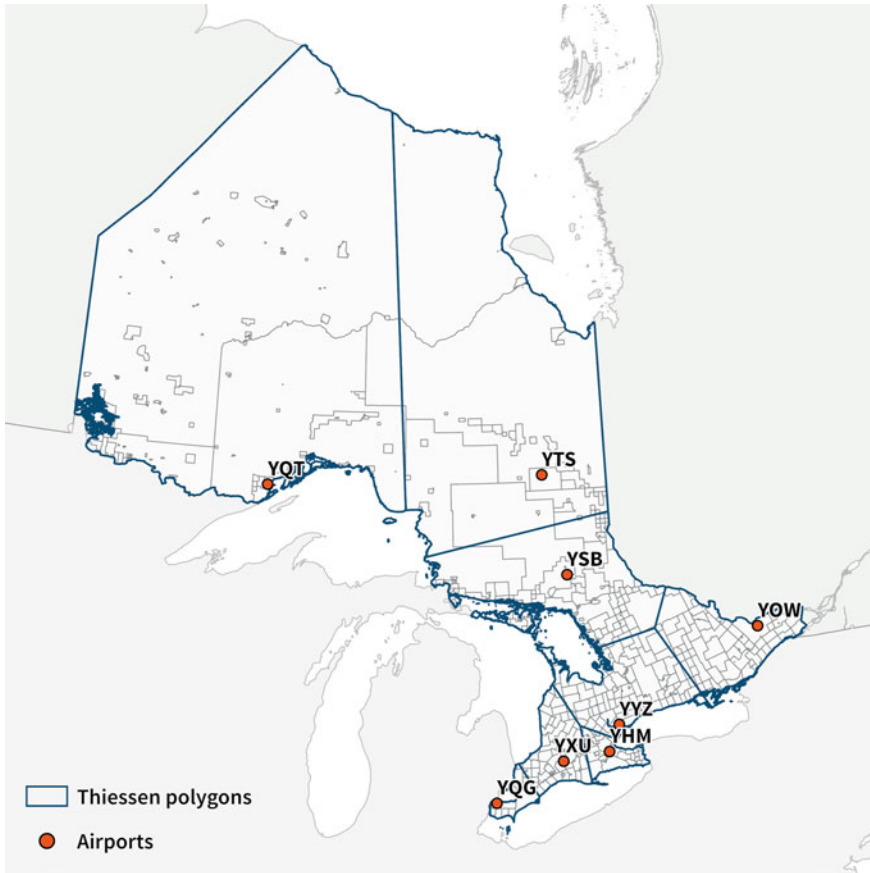


Fig. 8 Thiessen polygons for major Ontario airports overlaid on census subdivision boundaries

on flight volume, domestic passenger volume, and international passenger volume can be seen in the results for YTS. For this airport, subdivisions within close proximity have greater than 40% probability to use the airport based on total flight volume and on domestic flight volume. However, without reported international passenger volume, this airport disappears from the corresponding Huff model. Results for YXU and YOW followed a similar pattern, since they had fewer international passengers as compared to domestic passengers or total flights. In contrast, there was a clear increase in probability to use YYZ when attractiveness was based on international travel volume, specifically into northern Ontario. YYZ's catchment had an area of 981,143 km² when airport attractiveness was based on international passenger volume, as compared to 968,352 km² when it was based on domestic passenger volume (illustrated in Figs. 11 and 12).

Like Thiessen polygons, the Huff model produced large variation in the extent of each airport's catchment area. While for Thiessen polygons this variation was

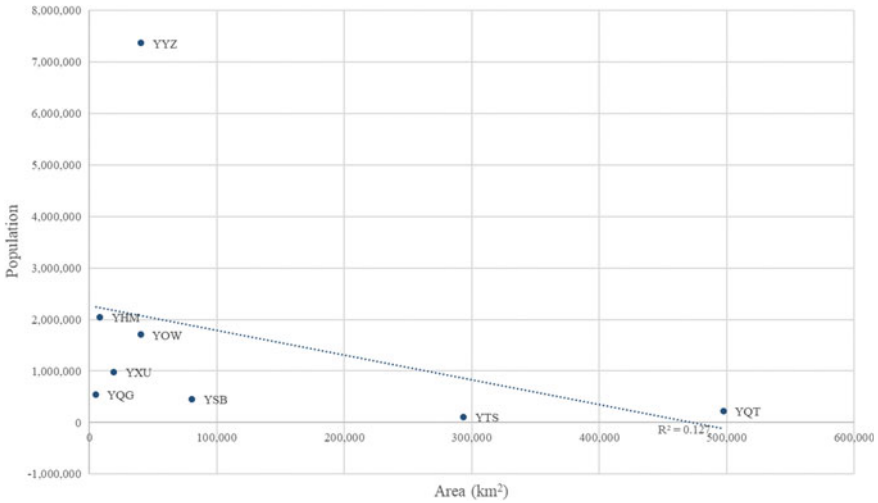


Fig. 9 Area and population for estimated airport catchment areas based on Thiessen polygons

the result of solely the distance between airports, variation produced by the Huff model was the result of both distance between airports and the attractiveness of each airport. Where YQT had the largest catchment area based on Thiessen polygons at 497,319 km², YYZ had the largest catchment areas based on all three variations of the Huff model at approximately 970,000 km² (Table 2).

This variation in area in turn impacted the catchment area populations with a strong, positive association between area and population based on the Huff model (Fig. 13). The Pearson’s correlation revealed that the relationship was significant for all three variations of results ($r = 0.982, p = 0.000$ for results based on passenger flights, $r = 0.997, p = 0.000$ for results based on domestic passengers, and $r = 0.998, p = 0.000$ for results based on international passengers. However, these statistics were likely impacted by YYZ being an outlier.

4 Discussion and Conclusion

This study illustrated that various market area delineation techniques produce notably different estimates of airport catchments. Differences existed in both the general shape and extent of the catchment areas as well as the land area and population associated with each. Moreover, area and population had varying relationships based on each technique. There were notable differences in the proportion of the study area belonging to each airport’s catchment based on the deterministic approaches (circular buffers, drive-time buffers, and Thiessen polygon) and the probabilistic approach (Huff model).

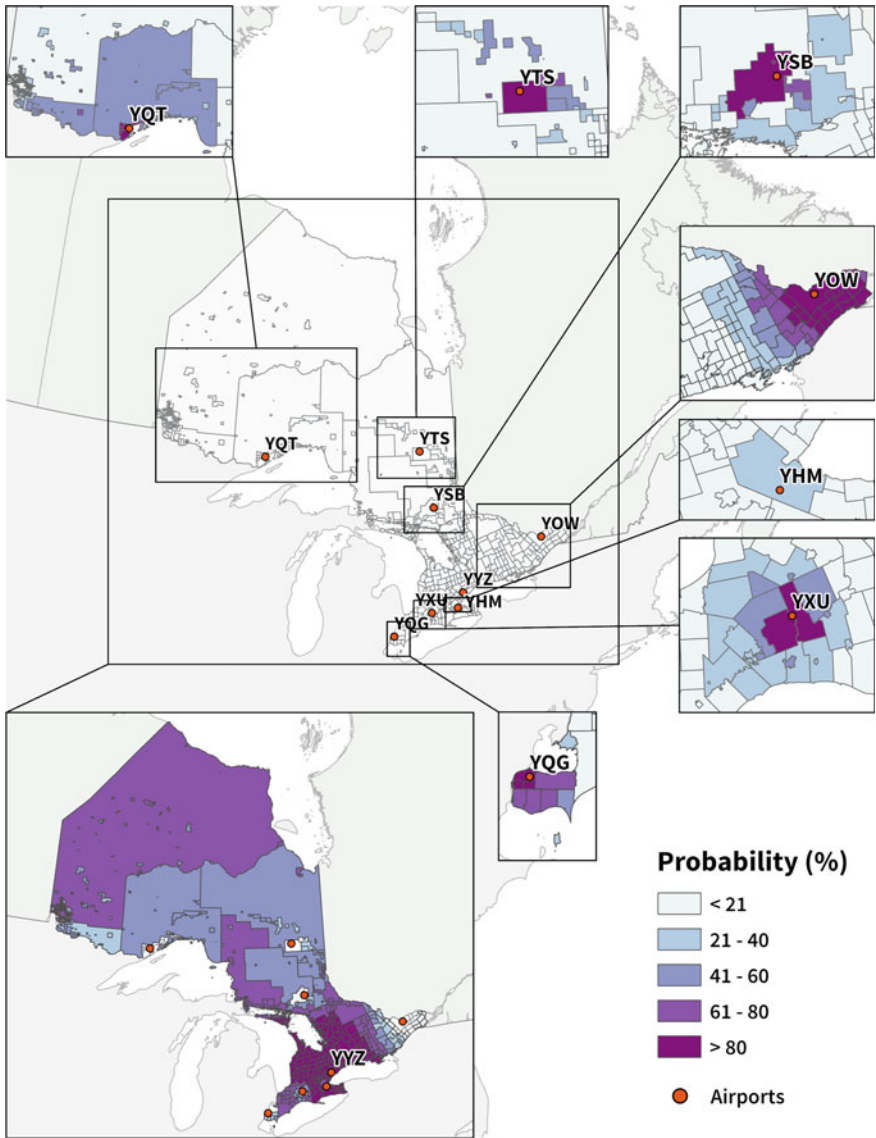


Fig. 10 Huff model results for airports with greater than zero total enplanements or deplanements, with airport attractiveness represented by total passenger flights, reported by census subdivision

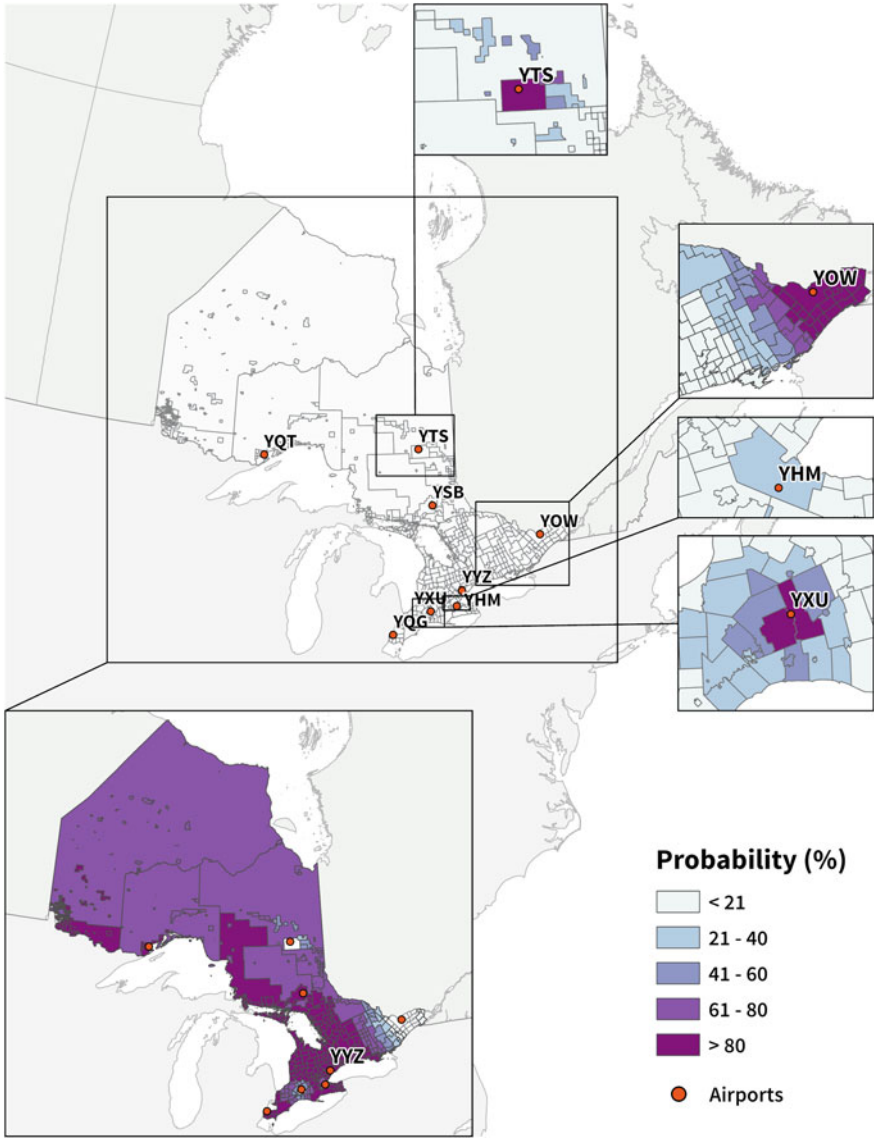


Fig. 11 Huff model results for airports with greater than zero total enplanements or deplanements, with airport attractiveness represented by total domestic passengers, reported by census subdivision (airports not appearing in an inset map had suppressed or unavailable data)

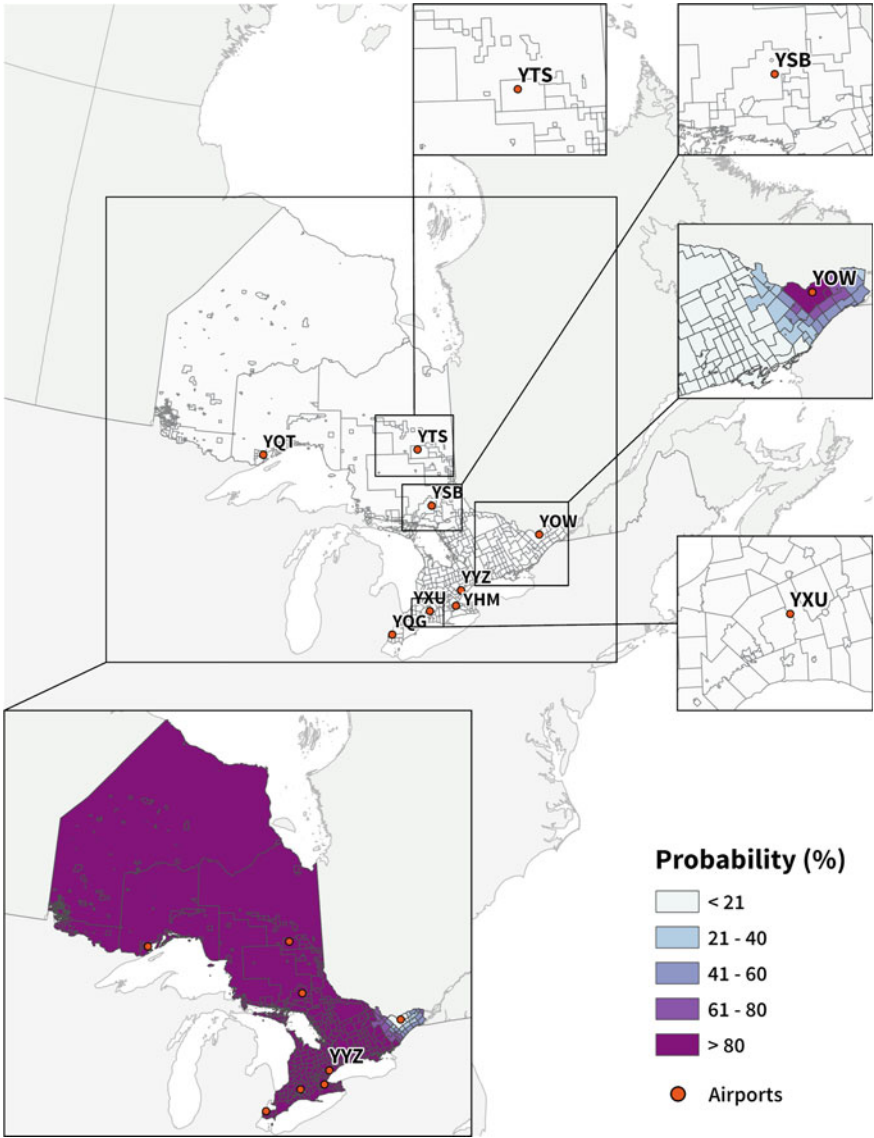


Fig. 12 Huff model results for airports with greater than zero total enplanements or deplanements, with airport attractiveness represented by total international passengers, reported by census subdivision (airports not appearing in an inset map had suppressed or unavailable data)

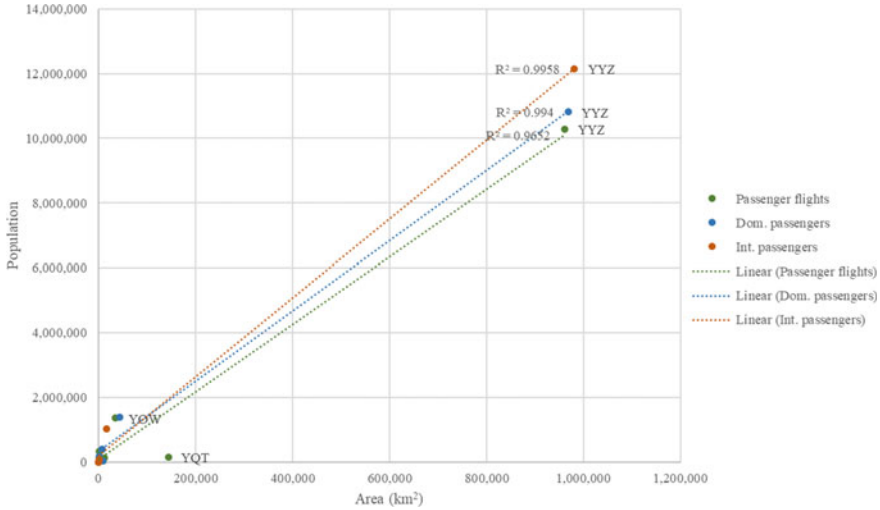



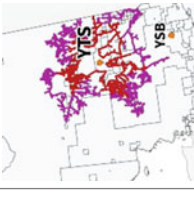

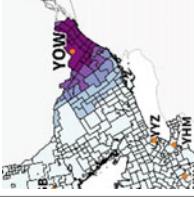
Fig. 13 Area and population for estimated airport catchment areas based on the Huff model, using total passenger flights, domestic passengers, and international passengers to represent airport attractiveness

The application envisioned is the forecasting of the spread of infectious diseases. Such analyses often must be conducted rapidly, with limited data. Therefore, the trade-off between efficiency of execution and validity of results must be considered when evaluating each technique. Table 3 provides a summary of the evaluation of each technique based on this trade-off. While this study was conducted based on the test case of Ontario’s eight major airports, it is expected that results are generalizable to other areas.

Of the techniques tested, the circular buffers and Thiessen polygon approaches would be the most efficient to execute, since their only data inputs are the airport locations. However, selecting a radius for the circular buffer technique might be difficult due to limited recommendations for an appropriate distance. This parameter may also depend on the specific public health at hand. There were no inputs parameters required for the Thiessen polygon technique. While this makes it efficient to execute, it is impossible to tailor the method to the use case and study area of concern.

The drive-time buffer also required only the single parameter of a defined drive-time cut-off, but required the additional dataset for the road network. If an analyst had a defined study area in which they knew their analyses would be conducted, this network could be pre-processed. However, if this technique had to be executed for locations globally, it would likely require additional resources to prepare a road network dataset. Moreover, infectious disease events may occur in remote or rural areas, or in developing countries, for which road network data often do not exist or are not readily available. Furthermore, road transportation via vehicle is not always the main method of transportation in remote areas. For example, during the recent outbreak of Ebola in Bikoro, Democratic Republic of Congo, the primary transporta-

Table 3 Characteristics of catchment area techniques and results

	Circular buffers	Drive-time buffers	Thiessen polygons	Huff model
Example				
Efficiency to execute	Very efficient	Efficient	Very efficient	Efficient
Data requirements	Minimal	Minimal, but road network required	Minimal	Fairly minimal, origin and attractiveness data required
Input selection	Buffer distance fairly arbitrary	Buffer distance fairly arbitrary	Cannot adjust inputs to suit use case	Parameterization required
Result type	Binary	Binary	Binary	Probabilistic
Airport weighting	Weighted equally	Weighted equally	Weighted equally	Weighting by attractiveness
Coverage of study area	Partial coverage	Partial coverage	Full coverage	Full coverage
Other advantages				Results appear most realistic
Other limitations		Driving not always main method of transportation	Does not allow for catchment overlap	Can extent unrealistically far (without distance cut-off)

tion route to the nearest major population centres was not by road but via the Congo River (WHO 2018).

While the three deterministic methods can be efficient, the results for our study area did not appear to be realistic in the context of the differences in attractiveness between airports; there are few airports in Ontario with high flight and passenger volumes. The distance-only deterministic methods might be more appropriate in cases where airports are equally distributed across a study area and are similarly attractive. Alternatively, for circular and drive-time buffer methods multiple buffer sizes could be applied. For example, a larger buffer could be applied for international airports or those with higher passenger/flight volumes, and a smaller buffer could be applied for domestic airports or those with lower passenger/flight volumes. Overall, deterministic methods could likely provide a general estimate of airport catchment areas in time-limited situations, though results would come with numerous limitations.

The probabilistic Huff model required additional inputs but was still relatively efficient to execute. In contrast to the deterministic methods, the Huff model estimated the greatest variation between catchment areas of the different airports. While this study did not incorporate validation of results based on observed data, the Huff model results appear to be reasonable considering the passenger and flight volume associated with each airport. It would be expected that people travel from much further to reach YYZ due to its high volume of flights, and because they might not be able to travel to their intended destination from smaller airports in Ontario. However, the Huff model estimated that subdivisions even at the northern border of Ontario were most likely to use YYZ. This is unrealistic and could be addressed by defining a distance cut-off in the model to restrict the choice set to airports within a reasonable distance. Additionally, the Huff model, as it was implemented in this study, does not account for the potential of flight connections. For example, an individual could take a domestic flight from YTS to YYZ, and then fly to an international destination from YYZ. Such itineraries could be represented in the attractiveness variable if more detailed data were available.

The characteristics of results based on each technique must also be considered based on the situation in which this approach would be implemented. Specifically, it is important to consider whether full coverage of the study area is required, i.e. whether all locations in the study area must be assigned to at least one airport catchment area. This may be needed if the spread of a local outbreak through outbound passengers was to be forecast. Only the Thiessen polygon and Huff model techniques meet this requirement, while the buffering methods do not cover the entire study area. However, if considering an inbound infected individual, catchment area results would not necessarily need to cover the entire study area to estimate the area within which the individual would likely travel after deplaning.

Considering the context of SARS, all methods would have assigned the City of Toronto to YYZ. This was expected, due to Toronto's close proximity to YYZ as well as the high volume of flights and passengers associated with YYZ. However, results based on each technique defined YYZ's catchment area differently in extent—affecting the area that an infected traveler who deplaned at YYZ might travel to. The catchment area for YYZ, in this case, could be considered at risk for SARS spread.

While the risk area based on Thiessen polygons was restricted to a small portion of southern Ontario, most of the province would have been included based on the Huff model. Resources might be focused on the defined risk area; therefore, it is important to select the most appropriate method, and apply appropriate data inputs, to best guide this distribution of resources. Considering YYZ's high flight and passenger volume compared to the other airports, it may be possible that its true catchment area covers a large portion of the province, as estimated by the Huff model. In this case, it could be useful to incorporate the population distribution within the catchment area to help guide resource distribution, as is common in the field of retail geography to estimate the spatial distribution of potential customers.

The results of this study were impacted by limitations in data access and conceptualization of air travel. In the airport locations and associated passenger dataset obtained from Statistics Canada, some data were suppressed and at least one important airport location was not included for unknown reasons. In future research, Billy Bishop Airport (YTZ) in Toronto, which serves approximately 2.8 million travelers annually (PortsToronto 2018), should be included to better understand its impact on the catchment area of Pearson International Airport (YYZ). Second, the comparison between results based on domestic and international passengers did not include travel to the U.S., since that is included in the "transborder" variable. Additional analysis could be conducted to analyze results based on travel to the U.S. specifically, or the variable could be merged with international passengers. Third, the open-source Huff model tool did not have an option to apply a distance cut-off to represent a maximum distance that an individual would be willing to travel. Without the use of a distance cut-off, Toronto's YYZ received high probability for individuals living in northern Ontario—in some cases over 1500 km away. The option to apply a distance cut-off would ensure a more realistic representation of travel to airports. Instead of travelling an unreasonably far distance, someone might choose to take a connecting flight from a local airport to a larger, more attractive airport from which to fly to their final destination. There was no obvious solution to including connecting flights in the modeling approaches, yet this would likely result in larger at-risk areas. Essentially, the catchment areas of larger airports would need to be extended to include the local catchments of smaller, connected airports, resulting in a network of airport catchment areas. Fourth, the methods did not consider airports outside of Ontario, such as US airports in Buffalo and Detroit just outside the Canadian border, which would likely reduce the estimated catchment sizes of some southern Ontario airports. Addressing this limitation would require expanding the study area to include parts of the United States. Fifth, we could not validate results with empirical data due to lack of access to such data. Validation is an essential next step. Primary data could be collected through surveys at airport locations or by collecting license plate information, or secondary data such as mobile phone data could be leveraged.

Infectious disease outbreaks can occur almost anywhere across the globe, and with high volumes of international air travel they can be imported to essentially any location. Thus, it is difficult to pre-define the study area that predictive analyses would need to be conducted for. With an understanding of the different characteristics of each market area delineation technique, it is important to focus on the requirements of

each particular use case. The airport catchment estimate should not only be as accurate as possible, but the technique must be manageable in situations where time and/or data are limited. This is often the case when there is an infectious disease outbreak, and analysis must be conducted rapidly to guide decision-making to respond and prevent further spread. In contrast to our case study, open-access airport and travel data would be more difficult to obtain globally, though options exist to purchase such datasets (e.g. from the International Air Transport Association). If a distance cut-off were incorporated to refine the results, the Huff model provides a balance between ease-of-execution and validity of results, so that catchment areas could be estimated rapidly and produce valid results to properly guide decision-makers when responding to infectious disease threats.

Acknowledgements Partial funding of this research from the Natural Sciences and Engineering Research Council of Canada is gratefully acknowledged.

References

- Augustyniak W, Olipra Ł (2014) The potential catchment area of Polish regional airports. *J Int Stud* 7(3):144–154. <https://doi.org/10.14254/2071-8330.2014/7-3/13>
- Başar G, Bhat C (2004) A parameterized consideration set model for airport choice: an application to the San Francisco Bay Area. *Trans Res Part B: Methodological* 38(10):889–904. <https://doi.org/10.1016/j.trb.2004.01.001>
- Bell DM (2004) Public health interventions and SARS spread, 2003. *Emerg Infect Dis* 10(11):1900
- Bilotkach V, Clougherty JA, Mueller J, Zhang A (2012) Regulation, privatization, and airport charges: panel data evidence from European airports. *J Reg Eco* 42(1):73–94. <https://doi.org/10.1007/s11149-011-9172-1>
- Bird BH, McElroy AK (2016) Rift Valley fever virus: unanswered questions. *Antivir Res* 132:274–280
- Bogoch II, Brady OJ, Kraemer MUG, German M, Creatore MI, Brent S, et al (2016a) Potential for Zika virus introduction and transmission in resource-limited countries in Africa and the Asia-Pacific region: a modelling study. *Lancet Infect Dis* 16(11):1237–1245
- Bogoch II, Brady OJ, Kraemer MUG, German M, Creatore MI, Kulkarni MA, et al (2016b) Anticipating the international spread of Zika virus from Brazil. *Lancet* 387(10016):335–336
- Boots B (1980) Weighting Thiessen polygons. *Econ Geogr* 56(3):248–259. <https://doi.org/10.2307/142716>
- Boots B, South R (1997) Modeling retail trade areas using higher-order, multiplicatively weighted Voronoi diagrams. *New York* 73(4):519–536
- Brent SE, Watts A, Cetron M, German M, Kraemer UG, Bogoch II, et al (2018) International travel between global urban centres vulnerable to yellow fever transmission. *Bull World Health Organ* 96(5):343–354B
- Centers for Disease Control and Prevention (2017) South Florida maps. Retrieved 5 Feb 2018, from <https://www.cdc.gov/zika/intheus/florida-maps.html>
- Cervero R, Round A, Goldman T, Wu K-L (1995) BART @ 20 series rail access modes and catchment areas for the BART system Robert Cervero Kang-Li Wu UCTC No. 307 The University of California Transportation Center University of California. BART @ 20 Series
- Debrezion G, Pels E, Rietveld P (2009) Modelling the joint access mode and railway station choice. *Transp Res Part E: Logist Transp Rev* 45(1):270–283

- Dolega L, Pavlis M, Singleton A (2016) Estimating attractiveness, hierarchy and catchment area extents for a national set of retail centre agglomerations. *J Retail Consum Serv* 28:78–90
- Fauci AS, Morens DM (2016) Zika virus in the Americas—yet another arbovirus threat. *N Engl J Med* 363(1):601–604
- Golnar AJ, Kading RC, Hamer GL (2016) Quantifying the potential pathways and locations of Rift Valley fever virus entry into the United States. *Transbound Emerg Dis* 65(1):85–95
- Government of Ontario (2018) About Ontario. Retrieved 2 Feb 2018, from <https://www.ontario.ca/page/about-ontario>
- Hatcher MJ, Dick JTA, Dunn AM (2012) Disease emergence and invasions. *Funct Ecol* 26(6):1275–1287. <https://doi.org/10.1111/j.1365-2435.2012.02031.x>
- Hernandez T, Lea T, Bermingham P (2004) What's in a trade area? Toronto
- Hess S, Polak JW (2005) Mixed logit modelling of airport choice in multi-airport regions. *J Air Trans Manage* 11(2):59–68. <https://doi.org/10.1016/j.jairtraman.2004.09.001>
- Huff DL (1963) A probabilistic analysis of shopping center trade areas. *Land Econ* 39(1):81–90. <https://doi.org/10.2307/3144521>
- Huff DL (2003) Parameter estimation in the Huff model. *ArcUser* 34–36. Retrieved from www.esri.com
- Huff DL, Black WC (1997) The Huff model in retrospect. *Appl Geogr Stud* 1(2):83–93
- Kilpatrick AM, Daszak P, Goodman SJ, Rogg H, Kramer LD, Cedeño V, Cunningham AA (2006) Predicting pathogen introduction: West Nile virus spread to Galápagos. *Conserv Biol* 20(4):1224–1231
- Leon S (2011) Airport choice modeling: empirical evidence from a non-hub airport. *J Trans Res Forum* 50(2):5–16. <https://doi.org/10.5399/osu/jtrf.50.2.2711>
- Levitt P, Jaworsky BN (2007) Transnational studies: past developments and future trends. *Ann Rev Sociol* 33:129–156
- Lieshout R (2012) Measuring the size of an airport's catchment area. *J Trans Geo* 25:27–34. <https://doi.org/10.1016/j.jtrangeo.2012.07.004>
- Lin T, Xia J, Robinson TP, Oлару D, Smith B, Taplin J, Cao B (2016) Enhanced Huff model for estimating Park and Ride (PnR) catchment areas in Perth, WA. *J Transp Geogr* 54:336–348
- Lounibos LP (2002) Invasions by insect vectors of human disease. *Annu Rev Entomol* 47:233–266
- McLay P, Reynolds-Feighan A (2006) Competition between airport terminals: the issues facing Dublin Airport. *Trans Res Part A: Policy and Practice* 40(2):181–203. <https://doi.org/10.1016/j.tra.2005.06.002>
- Muller MP, Richardson SE, McGeer A, Dresser L, Raboud J, Mazzulli T, Canadian SARS Research Network, et al (2006) Early diagnosis of SARS: Lessons from the Toronto SARS outbreak. *Eur J Clin Microbiol Infect Dis* 25(4):230–237. <https://doi.org/10.1007/s10096-006-0127-x>
- Ontario Ministry of Finance (2017) 2016 census highlights: factsheet 8. Retrieved from <https://www.fin.gov.on.ca/en/economy/demographics/census/cenhi16-8.html>
- Ontario Ministry of Natural Resources (2010) Ontario road network: segment with address, captured February 2010. Retrieved from <https://www.ontario.ca/data/ontario-road-network-segment-address>
- PortsToronto (2018) Billy Bishop Toronto City Airport. Retrieved from <https://www.billybishopairport.com/>
- Powers AM (2015) Risks to the Americas associated with the continued expansion of chikungunya virus. *J Gen Virol* 96(1):1–5. <https://doi.org/10.1099/vir.0.070136-0>
- Reilly WJ (1931) *The law of retail gravitation*. University of California, New York
- Sanko N, Shoji K (2009) Analysis on the structural characteristics of the station catchment area in Japan. In: 11th conference on competition and ownership in land passenger transport
- Statistics Canada (2016a) Air passenger traffic and flights—Table 401-0044. Retrieved from <http://www5.statcan.gc.ca/cansim/a26?lang=eng&id=4010044>
- Statistics Canada (2016b). Boundary files. Retrieved from <http://www12.statcan.gc.ca/census-recensement/2011/geo/bound-limit/bound-limit-eng.cfm>

- Statistics Canada (2016c) Census of population. Retrieved from <http://www12.statcan.gc.ca/census-recensement/2016/as-sa/98-200-x/2016001/98-200-x2016001-eng.cfm>
- Summers A (2013) Pandemic flu: lessons from the Toronto SARS outbreak. *Emerg Nurse* 17(5):16–19
- Suzuki Y (2007) Modeling and testing the “two-step” decision process of travelers in airport and airline choices. *Trans Res Part E: LogTrans Rev* 43(1):1–20. <https://doi.org/10.1016/j.tre.2005.05.005>
- Tatem AJ, Hay SI, Rogers DJ (2006) Global traffic and disease vector dispersal. *Proc Natl Acad Sci* 103(16):6242–6247
- The SARS Commission (2006) Spring of fear. SARS commission final report, vol. 2. Toronto
- Upchurch C, Kuby M, Zoldak M, Barranda A (2004) Using GIS to generate mutually exclusive service areas linking travel on and off a network. *J Trans Geo* 12(1):23–33. <https://doi.org/10.1016/j.jtrangeo.2003.10.001>
- Wang F (2000) Modeling commuting patterns in Chicago in a GIS environment: a job accessibility perspective. *Prof Geo* 52(1):120–133. <https://doi.org/10.1111/0033-0124.00210>
- WHO (2018) Ebola virus disease—democratic Republic of the Congo. Retrieved from <http://www.who.int/csr/don/10-may-2018-ebola-drc/en/>
- Wittman MD (2014) An assessment of air service accessibility in U.S. metropolitan regions, 2007–2012. Cambridge.
- Yamada I (2016) Thiessen polygons. *Int Encycl Geogr People Earth Environ Technol* 1–6

Reflective Practice: Lessons Learnt by Using Board Games as a Design Tool for Location-Based Games



Catherine Jones and Konstantinos Papangelis

Abstract Location-based gaming (LBG) apps present many challenges to the design process. They have very different requirements compared to games that are aspatial in nature. They take place in the wild and this brings unique challenges to the practicalities of their design. There is a need to balance the core game play with the spatial requirements of location-aware technologies as well as considering the overall theme and objectives of the game together with the motivations and behaviours of players. We reflect upon this balancing act and explore an approach to creative paper prototyping through the medium of board games to co-design LBG requirements. We examine two case studies of location-based games with different goals. The first case study discusses the CrossCult Pilot 4 app built to trigger reflection on historical stories through thoughtful play. Whilst the second case study uses the City Conquerer app designed and played in Suzhou, China with a view to exploring notions of territoriality. The paper considers how spatial, social and interaction metaphors are used to simulate location-based games in a board game and discusses the lessons learned when transforming the paper game into a digital prototype. It forms part of a thinking by doing approach. By comparing the board games to the technical counterparts, we consider how effective are the features and activities implemented in the technology prototypes. We propose a set of 11 design constraints that developers must be mindful of when transitioning from paper to digital prototypes.

Keywords Location-based games · Design · Prototypes · Board games · Game design · Urban games · Smart cities · Playable cities

C. Jones (✉)

Department of Geography and Spatial Planning, University of Luxembourg, Luxembourg City, Luxembourg

e-mail: catherine.jones@uni.lu

K. Papangelis

Department of Computer Science and Software Engineering, Xi'an Jiaotong-Liverpool University, Suzhou, People's Republic of China

e-mail: K.papangelis@xjtlu.edu.cn

© Springer Nature Switzerland AG 2020

P. Kyriakidis et al. (eds.), *Geospatial Technologies for Local and Regional Development*, Lecture Notes in Geoinformation and Cartography, https://doi.org/10.1007/978-3-030-14745-7_16

291

1 Introduction

We often take our cities for granted. Passing through them as we go about our daily routine as we travel to and from work. So busy engrossed with our daily activities, we stop noticing, exploring and interacting with that which is, and those whom are around us. With the advent of location-enabled mobile phones, we are witnessing the emergence of new playful forms of interaction and participation between and within citizens and the city. Which changes the activity and materiality of the city and supports different types of encounters. Take for example here in Luxembourg at the height of the Pokémon Go phenomenon, the city parks were transformed from the domain of the family and in particular, children to that of couples and individuals, they became crowded with players actively using the space to catch Pokémon. The geolocated game on the mobile device transformed and disrupted the normative materiality and activity of the place, into a place of hybrid play. Seamlessly intertwining the digital and the physical worlds bringing a blended approach to the city.

The embedded nature of locative technology within our smart phones has paved the way for a broad range of research in this field (de Souza e Silva and Frith 2012; Saker and Evans 2016). We are witnessing a growing market in location-based games, which engage millions of players, take PokémonGo, Geocaching and Ingress as examples. In addition to these global successes there are also many niche apps that still successfully engage outdoor play and urban discovery, albeit at a different scale of engagement. GPS and mapping technologies provide users with playful experience where the city is transformed into a living board game or playground. If designed appropriately location-based games stimulate playful urban discovery, interaction, sociability, information retrieval and the like (Ahlqvist and Schlieder 2018; Nijholt 2017).

1.1 *Board Games and the Design of Location-Based Games*

To summarize, the studies noted above illustrate that the City is an intrinsic component of a geo-located game (Evans and Saker 2017) but often it forms a passive component because the game play simply uses the city as a container of activity but not much more. To provide a fully immersive and embodied experience that shapes our playful interactions we need to design location-based games that facilitate interaction and dialogue between the city and its citizens. If carefully planned and developed, these transformative forms of city-based engagement could be encouraged and achieved using geolocated gaming apps. Such technology ecosystems need novel techniques and methods that enable them to be (co)designed creatively and efficiently. This is because location-based gaming apps present many challenges to the design process. They have very different requirements compared to games that are aspatial in their nature. They take place in the wild and this brings unique chal-

lenges to the practicalities of their design. There is a need to balance the core game play with the social-spatial requirements of being outdoors together with the overall theme and objectives of the game with the motivations of players. There must also be consideration for notions of place attachment and territoriality that manifest because of the place-based ties players have.

There are many existing methodologies for designing and evaluating online games or websites that have been tried and tested ranging from cardboard prototyping (Ehn and Kyng 1992) to Wizard of Oz prototyping (Li et al. 2004). Thus, it seems natural that board games can aid in the planning of location-based games, especially since they are intuitive, playful and sociable and incorporate many elements that are used in location-aware games (proximity, movement, social interaction). It seems sensible that common metaphors such as those used in board games should become part of the location-based game designers' toolkit. Indeed, the Carcassonne game was used as the starting point for the development of a geolocated game discussed by Matyas et al. (2008). In the literature, a few studies integrate board games into the iterative design process of location-based games. These studies take inspiration from existing board games (Mateos et al. 2016; Schlieder et al. 2006) or custom design a game based on their unique scenario (Jones et al. 2017a; Marins et al. 2011). They build upon notions of paper prototyping incorporating playfulness and sociability to foster a co-design process.

1.2 Methodological Overview

We reflect upon this balancing act and explore an approach to creative paper prototyping using the medium of board games to co-design game requirements. We examine two case studies with very different goals. The first case study in the spotlight, discusses the CrossCult Pilot 4 app, designed to provoke reflection and reinterpretation on historical topics through thoughtful play, whilst the second case study uses the City Conquerer app that was designed and played in Suzhou, China that explored notions of territoriality. We use reflective practice to look back on the paper and technology prototypes to form part of a process of thinking as doing. Thus, the discussion reflects upon our lessons learned that result from our thinking and observations of the process that started with board games and comparing them to the technical implementation that were subsequently evaluated. We identify strengths and weakness of the use of board games for this design process.

2 Case Study 1: CrossCult PILOT 4—Designing Games for Serendipitous Urban Discovery and Personal Reflection

2.1 App Overview

The app is motivated by the desire to foster reflection and reinterpretation through location-based experiences by encouraging participants to encounter different types of historical stories, which relate to current social challenges such as migration. The app is best described as a serendipitous urban discovery game. It triggers reflection on the cultural heritage of cities and their public history by encouraging users to engage in playful encounters. This supports an intellectual journey. Through geolocated play with the city, its history and its people are brought into our consciousness. As player’s walk through the city they choose which points of interest (POIs) to discover based on where they are and where they are going (Fig. 1). They can also have the phone in their pocket and it will vibrate if they are passing through a POI.

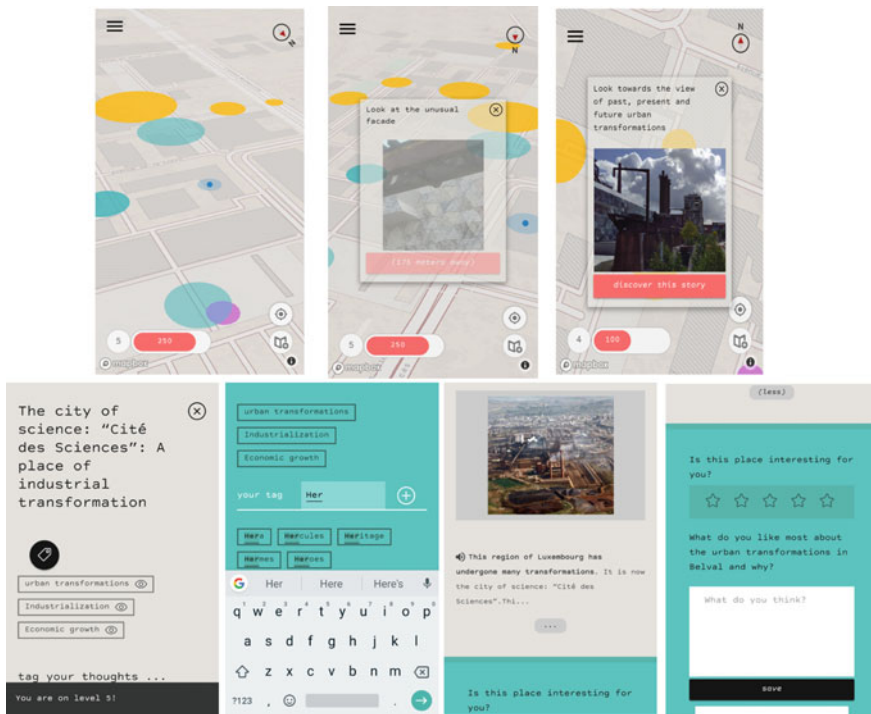


Fig. 1 Screenshots showing CrossCult Pilot4 App. (top left) birds eye map, (top middle) navigational clue, (top right) story discovery, (bottom left) story window with title and tags, (bottom middle) story media and text (bottom right) rating and answering a question



Fig. 2 Board game set up (left) and game play (right)

POIs are comprised of navigational clues (a title and an image) which indicate to players the story locations. The text of a navigational clue hints towards the topic of the story. The POIs are symbolised on the map as yellow or purple circles—yellow for curated stories and purple for player-contributed personal stories, see Fig. 2. As soon as the user’s physical location (blue dot) is located within the yellow or purple circle, users can open, read and interact with the story.

The stories discovered can have either a direct connection to the POI or a metaphorical connection to the location. Each story is comprised of short self-contained texts together with a contextual image. As part of the game, players can win points, move through game levels, unlock functionality and earn achievement badges for completing different tasks including answering questions based on the stories. Comments on questions are marked as pending until they have been moderated via a companion app. The backbone of the app is the location tracking system and its ability to function as a sensor capturing the player’s location and deliver information accordingly.

2.2 Board Game Design Approach and Discussion

We used an initial scenario to inform the design of a board game that became a tool for both paper prototyping and co-design. The final design results have been previously described (Jones et al. 2017b), but in essence play tests were conducted with 16 participants (from the author’s institute), across three play sessions. The game attempted to mimic how players would move, discover and interact with stories by building upon existing intuitive representations of movement and play that established in board games. We used a game master to organize the play sessions, their role was to explain the rules, log the points on the scoreboard and manage the content interactions (making it available when players reached points of interest).

Urban movement and playspace: The act of movement through the city and discovering POIS, central to the mobile game, was simulated using a combination of

features. That were typical of many board games in which players move around a board depicting an imagined landscape. By rolling a dice they could move their *playing piece* by counting the number of steps and stopping when they reach a landmark (POI). For our play space, we replicated this movement metaphor using a paper map (Luxembourg). Onto which we marked the real locations of the POIS (aligned to their GPS coordinates) and then drew paths between them. Stepping-stones that represented nodes on the street network were marked at junctions in the real street network. They represented the choice of turning and changing direction midway through long streets—they marked the effort required to move along long streets (see Fig. 2). The stepping-stones together with the score from the roll of the die restricted how far you could ‘walk’ in an attempt to simulate the physical effort of moving in a city. We replicated autonomy of movement by enabling individual players to begin the game at any location marked and then with their roll of the die they were free to move towards any POI of their choosing. The design supports serendipitous discovery of stories as opposed to traditional linear tours through a city that have a designated start and end-point, players could ‘walk’ towards any POI they wanted and change direction at any time.

Situated story activation: In board games, players land on special tiles where players collect or pick up playing/event cards (think of monopoly and the chance card). These cards require the participant to take an action (go straight to jail, pass go or collect \$200). We took this idea of the event playing card to replicate the screen of the mobile (creating a postcard size booklet), simulating the access to content that could be discovered. Story discovery was activated when a user reached a POI icon on the map. When players found themselves in the correct location, they were given the option to open and interact with the story booklet or they could move on if they were not interested. Players were under no obligations to interact with content unless motivated to do so.

Context awareness: We were keen to design the game to provide a mechanism where players were not obligated to spend the entire game looking at their mobile phone. Context awareness was an important concept. The phone automatically notifies you when are in the vicinity of a POI, using haptic feedback if the app was open in the background. To evaluate this type of feature in the board game, a game master played a sound when a player approached certain POIs.

Reading stories and reflecting: In board games playing/event cards, have actions, rewards, or penalties that contribute to the winning conditions of the game. Thus, we adapted the concept of event cards to create a postcard booklet—which you could open when you arrived at the POI. Players could not read the hidden story until this point was reached. We also used the small a5 size booklets as an imitation of the mobile screen with each page more or less representing a scroll down on the screen. Page 1 showed the navigational clue picture and a more detailed map, the second contained a historical picture and the story whilst the third page focused on the interactions of tagging and responding to a question (see Fig. 3).

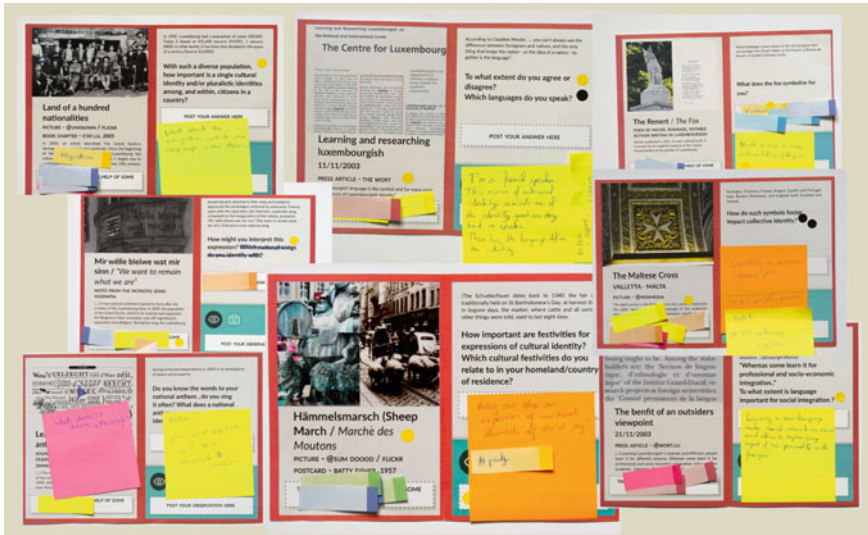


Fig. 3 Postcard booklet with participant interactions

Social interaction and player contribution: A key aspect of board games is their ability to create shared social experiences. They come in many different formats which include designs to create competition between players or those that are based on consensus where players work together to achieve a common goal. We needed to design elements of the game that would collect user-generated content and encourage other users to view the contributions of others. When a player landed on a story they could choose from a number of interactions: (1) tagging a thought on a story, (2) rating your reaction to a place, a story or content or question (3) answering a question designed to stimulate reflection. We used post-it notes together with agree/disagree stickers (see Fig. 3) and all interactions with the booklet won players points.

Geolocated reflection(s) and personalized stories: We had one further game mechanic designed to enable users to share their story at location of their choosing. If players rolled a 1, they could spend that turn to pick a *treasure card*. In the first iteration of the board game players would write on blank treasure or read treasure cards provided by others. In subsequent iterations (there were 3 in total) story cards used intuitive symbols to prompt a personal memory that could be pinned to any location. This interaction simulated co-creation of user-created points of interest.

2.3 Lessons Learnt

With the Pilot 4 app designed in beta version and undergoing an extensive evaluation, it is time to reflect back on the design process. Paper prototyping has long been recognized as a useful tool for human computer interaction design (Snyder 2003)

and so it seems only natural that in the design of location-based games we combine the notion of paper prototyping and board games—there are so many that simulate interaction of being in the city. We were able to evaluate and co-design a number of different aspects including the (a) cultural content presentation and quality (historical objects and narratives), (b) the participative activities (reflections, interactions and user memories), (c) the winning conditions, and (d) the game design. The implementation of the digital app led to a series of evaluation experiments results of which are described in a publication (Jones et al. 2019).

2.3.1 2D Paper Maps for Complex Topography

We used a 2D paper map as game board. With such a flat object, it is difficult to replicate the presence of urban barriers such as landscape topography. In the board game, people had no worries moving freely. In reality Luxembourg is undulating, it lacks spatial continuity and has natural barriers (a large valley) which requires extra physical effort. These barriers are challenging to replicate on paper and the app requires game mechanics to motivate you to go to inaccessible places.

2.3.2 Sociability and Competitiveness

Social interactions are often a driver of the gameplay and help to stoke competition amongst the players. In board games players are able to see each other positions and observe their paths of movements and this helped to motivate players. This is where there is a slight divergence with the technology prototype and the board game design process. We felt it inappropriate to reveal the locations where other players had been—due to data privacy and ethical issues of mediated co-proximity and geostalking (Licoppe and Inada 2009).

2.3.3 Modelling Asynchronous Interactions

The mobile app of Pilot 4 app was always conceptualized as an asynchronous game, players interacted by leaving their reflections which other players discovered. In the board game, this asynchronicity was simulated by keeping the story booklets closed and only allowing players to interact with them when it was their turn and when they arrived at the correct location. To replicate the true form of asynchronous game, on a board game we would have had to hold separate game play sessions with each individual participant playing on their own—which we chose not to do so for logistical, social and playful reasons. We made a compromise.

2.3.4 Bird’s Eye View of the Map

In the board game the players had a birds eye view of the locations of the POIS and their distribution in the city. We failed to carry this perspective through into the technical implementations. During an early prototype, we restricted the zoom level, preventing players from seeing this bird’s eye view of all the locations. The team thought this would be a more playful interaction. An early round of testing with an external participant indicated the value and importance of players that this view had. In later experiments we observed participants using the bird’s eye view to plan their routes and decide where to go. Mirroring the same behavior patterns that was used in the board game, indicating its use and benefit to participants.

3 Case Study 2, CityConqueror—Designing Games for Exploring Territoriality

3.1 App Overview

CityConqueror was inspired by the board game ‘Risk’. In this board game a player conquer countries on a world map, deploying units to defend and attack countries (Fig. 4 left). In the game CityConqueror, players can conquer territories in their physical location, deploy units to defend their territories and attack those of other players that are nearby. It was designed for the neighbourhood scale.

In the technology prototype when conquering a territory, the player gives it a name that is then visible to other players. They can deploy units to defend the territory and hide a treasure in it. Territories are conquered and plotted on a map of the “real” urban terrain, showing the player’s location. The map is covered by the *Fog of War* similar

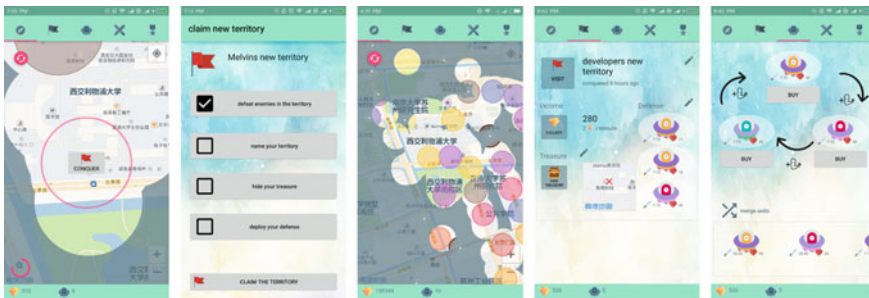


Fig. 4 (left) Fog of War: player conquer territory based on h current physical location. (middle left) To conquer a territory, the player gives it a name, hides a treasure and deploys units. (middle) Territories are plotted on the map. (middle right) A territory generates resources. (right) There three different types of units

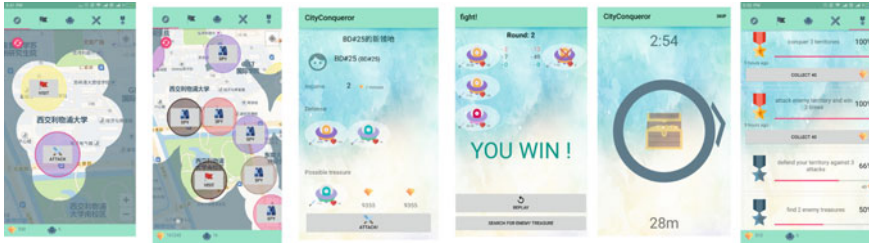


Fig. 5 (left) To conquer a territory the player must defeat all enemy territories (the pink ring in this figure). (middle left) Players can spy on enemy territories from outside their range. (middle left right) When spying on a territory, players can see who conquered the territory, the name of the territory, and other information (middle right left) The result of a fight. (middle right) After winning the player can try to find the treasure by getting to the indicated position within the time limit. (right) player can complete achievements

to other popular games that deal with territorial conquests, such as *Age of Empires*.¹ Players uncover the Fog of War by physically exploring the urban space. When a player visits a physical space they remove the Fog, they are able to see the enemy territories. To motivate exploration of the player's physical surroundings, hotspots in the undiscovered spaces are marked on the map, indicating the location of enemy territories in the Fog of War (Fig. 5 left). Territories generate resources over time, which can be collected to buy units to defend their own territories or attack enemy territories (Fig. 4 left).

To conquer a new territory, players must defeat all enemy territories in their range (Fig. 4 middle). If a player attacks an enemy territory and wins, they have the chance to find the hidden treasure. Searching for a treasure is a mini-game. The player is given a compass that points in the direction of the treasure, told how far away it is and the has three minutes to find it (Fig. 5 middle right). The objective of *CityConqueror* is to: (1) claim as many territories as possible, (2) generate income, (3) defend territories against attacks from others or (4) to attack others and (5) conquer their territories (Fig. 5). Thus, the game experience of one player is highly dependent on the actions and interactions of other players.

3.2 Design Approach and Discussion

We began by brainstorming. This resulted in an approach that involved developing the core functionality of *CityConqueror*, and testing various mechanics in the wild. This seemed like a good idea but once we started thinking of the practicalities, it became immediately apparent that it would take time until there was a viable technical prototype. Therefore, we changed tactic. At the same time as developing the game

¹<https://www.ageofempires.com/>.

engine we employed Schlieder's (Schlieder et al. 2006) approach and designed a board game to explore the game experience and mechanics.

Play space: We developed the play space by mapping the city of Suzhou in a hexagonal 60×60 in. game map using hexographer.²

Urban movement: To create synchronous movement and asynchronous gameplay, we used the *turn tokens* mechanic of *Camelot*.³ In brief, the "the turn token" comprised of 2 tokens that were passed in opposite directions. The players with the token can move 1 hex or complete an action. Once the turn is over, they pass *token 1* to the person in their left or *token 2* to the person to their right. This allowed players to take turns quickly in an asynchronous manner. It forced a fast and intense game.

Interacting with territories: Individuals were actively encouraged to pursue a goal by interacting with each other and the play space. We employed goals (e.g. capture a territory) and winning conditions similar to *Risk*.⁴ *Tokens (game pieces)* denoted *interactions* with territories and players used tokens to simulate interaction with different territories. All players started with 30 area tokens. It cost 1 area token to conquer an unoccupied area (removing the fog of war), attacking territories cost 2 tokens whilst defending your area did not cost tokens.

Conquering territories: The conquering and defending of territories was simulated using a simple *rock-paper-scissor-lizard-spock* game. In this variation *Spock* smashes *scissors* and vaporizes *rock*; he is poisoned by *lizard* and disproven by *paper*. *Lizard* poisons *Spock* and eats *paper*; which is crushed by *rock* and decapitated by *scissors*. *Spock* is signified with the Star Trek *Vulcan* Salute, while for "lizard" you form the hand into a sock-puppet-like mouth. This variation led more often to a win rather than a draw.

3.3 Reflecting on the Design Process—Lessons Learnt

Once we were happy with the game and the rules we organized a series of play sessions with 12 groups of 8 players each (79 male and 97 female). They lasted approximately 60 min per group and involved (1) a brief intro presentation to the game, (2) playing the game and (3) an informal group discussions on how to improve the board game and how to make a mobile game out of it. The results of which are describe in (Papangelis et al. 2017a, b). By reflecting on the observations we identify a further group of lessons learnt.

²<http://www.hexographer.com/>.

³<https://boardgamegeek.com/boardgame/14812/camelot>.

⁴<https://www.hasbro.com/en-us/product/risk-game:2C7C6F52-5056-9047-F5DD-EB8AC273BA4C>.

3.3.1 Social Interaction and Combat Systems

During the play test it was clear that the combat system, whilst fun, would not transfer easily to the mobile device. We had to develop a ‘turn-based’ mechanic used in popular pen and paper role-play games such as Dungeons and Dragons.⁵ A player in a turn-based game is allowed a period of analysis before committing to a game action, ensuring a separation between the game flow and the thinking process, which is thought to lead to better choices.

3.3.2 A Light-Weight Digital Game Layer and Story Plots

The board game constructed an imagined layer of the city, which was transposed to the digital prototype. This created a lightweight and thin virtual world. It offered a ‘real feel’ and enabled players to have the illusion of playing in the real world. The use of game mechanics blurred the borders and interpretations of ordinary space and play for the players because the game was based on an ‘*actual*’ map of the city.

We chose not to include a plot or characters in the game, we wanted to create a thin hybrid reality to blur the borders and interpretations of ordinary space and play for the players. Such an approach is based upon Flanagan’s view (Flanagan 2013) that locative media and pervasive games should refer to, and not appropriate spaces, therefore not divorce them from their meaning, history, and significance.

3.3.3 Personal Perceptions and Use Urban Space

Players used different strategies to conquer territory in the board game compared to the digital prototype. In play tests with the app, players used established strategies based upon personal attachments and perceptions of place to influence where they went. For example, players took control of familiar areas and then expanded by conquering neighboring territories. This behavior was less dominant in the board game, so game mechanics must be mindful of this. Players’ perceptions of space and place such as the way meaning, history, and the significance of place are appropriated (e.g. historic city centre) are difficult to simulate in the board game.

3.3.4 Hybrid Nature of the Magic Circle

In game design, the magic circle describes when players start using games (location-based) they implicitly agree to some interactions happening in the virtual world, these virtual interactions stand consciously outside of everyday life (Huizinga 1955). The magic circle helps to transport the players into hybrid space where physical and digital play meet. We observed a difference between the imagined spaces created in

⁵<http://dnd.wizards.com/> accessed 12/21/2016.

the board game to the technology prototype. In the city players were neither “in” or “out” of the magic circle, but rather continuously hoping between both. In the board game it was easier to be drawn into the magic circle. LBG should not draw the players into an entirely imaginary world in the way board/video games do, rather they should encourage play with real spaces (Benford et al. 2006) by a focusing on the interplay of the gameplay, content and physical locations. This could be achieved by (1) incorporating elements of the ordinary world into the gameplay, and (2) through designers taking into account both physical attributes of cities and digital media.

4 Reflective Discussion Constraints of Using Board Games of LBG Design

The use of board games in both case studies provided a tool to encourage dialogue between project members and participants and helped the evaluation of early game scenarios. They also assisted in the co-design of the requirements, aiding the identification of refinements and new ideas through collaborative participation, helping to stimulate imagination and enthusiasm in a playful manner. It was a fun way to engage people outside of the project in the design process. Researchers observed the players interacting with their scenario from the outside-in, noting what worked or not as they watched the gameplay. Players were also able to comment, contribute and thus become collaborators with the research as part of the creative process of the project. This provided an inside-out perspective resulting from involvement in the participation. The board games prototypes supported the process of thinking by doing and reflective practice.

By examining the two case studies, which focus on different types of LBG, we were better able to understand the advantages and limitations that result. The case studies illustrate that even though board games can be a useful tool in the arsenal of an LBG design and co-design, they cannot fully emulate all the necessary mechanics for an engaging gameplay. This is because of the limitations of modelling the complexity of the real world in a constructed environment of a board game. Thus, we propose a set of constraints that one must be mindful of when is transitioning from a board game paper prototype to the technical prototyping, see Table 1.

One of the most significant differences we observed were the manifestation of local attachments that emerge during real location-based game play in areas that are familiar to us. Any such playfulness should be designed to mediate and moderate the inherent local ties and place-based bonds that we have. Particularly if the game play is to take place in areas that are familiar to us, we are more likely to be comfortable playing in neighborhoods that we feel connected to, so players are likely to feel an outsider in unfamiliar parts. These feelings are less likely to emerge in a paper prototype. When one adapts a board game to an LBG it is necessary to pay attention to the person-to-person and person-to-place aspects of place attachment and sense of place. As these relationships can result in expressions of attachment that are played out because there is a sense of belonging to one place and not to another. It is our

Table 1 Constraints of using board games in the process of location-based game design

ID	Description	Category
1	In the board game players tend to be engaged in playful and competitive interactions that are beyond the rule of the LBG	Playful interactions
2	The magic circle of the board game is more pronounced as the players are strongly embedded in creative play	Playful interactions
3	LBG facilitate asynchronous and anonymous interactions with strangers—less likely in a board game as players are together	Playful interactions
4	Effort and fatigue required to go to more distant parts of the city need to be designed for in LBG technology prototypes	Discovery
5	Autonomy of movement and choice of content should be facilitated	Personalised discovery
6	Depending upon the LBG game person-to-person and person-to-place attachments emerge which are not evident in the board game	Manifestation of local attachments
7	People are comfortable in certain real locations, potentialities of power emerge (such as expressions of territoriality, place attachment and avoidance of locations)	Manifestation of local attachments
8	A sense of commonality and relationships can be fostered in players of LBGs which are absent in the board game	Social-spatial bonds
9	Cities can actively contribute to the co-creation process of local knowledge in LBGs because they offer a multisensory and immersive environment	Co-creation of local knowledge
10	LBG create a collectivity of movement beyond the everyday	(Re)familiarisation
11	LBG are prone to privacy and surveillance issues (e.g. location tracking of real-time movements) which is not an issue with board games	Ethics and privacy concerns

belief that the management of these ties should be designed to either avoid/enhance these expressions of territoriality and attachment and/or to encourage the discovery on unexpected, surprising locations. This depends upon the underlying goals and objectives. We suggest designing features and gameplay that support such goals.

Firstly, we suggest the implementation of features to the articulation of sense of identity and place attachment as part of the gameplay. Secondly, designs should consider the power of the player-place attachments and be mindful to reinforce or challenge them. This could be implemented using game missions, which require participants to go to unfamiliar parts of the city and to complete participative situational-based tasks motivated using incentives and rewards as part of the game framework.

Moreover, person-to-person and person-to-place ties to places are forged by groups of people that have relationships that embody a sense of familiarity or commonality (Chamberlain et al. 2017) with each other and with places (social-spatial bonds). These again are not explicit in board games since the play-space is static and less relatable—it is an abstraction. An imagined representation of reality. One that is simply a model, a more immaterial version of reality. To encourage sociability that is present in board games we suggest functions that promote the creation of communities are taken into account. These may be user defined based on preferences (e.g. students of Xi'an Jiaotong-Liverpool University) or system created based on the place attachments, a particular group of users. Since these interactions are hybrid, they exist both in the virtual and the physical realm we suggest their design incorporates both in-place and out-of-place interactions. This might encourage self-awareness in their personal place-based meanings and consequently payers become aware of their influence on their personal identity (Chorley et al. 2015). Avatars can be also integrated into the design to build up digital identities that individuals/groups can identify with.

5 Conclusion

Board games are fun and playful and are a valuable tool for the co-design process of location based games. They provide many advantages since they are easy to facilitate and are more playful than chauffeured pencil-paper prototyping methods. Chauffeured pencil-paper prototyping requires the designer and the players to be in the city and interact with pieces of paper that simulate the interface. They involve the designer describing to the user and manually demonstrating how the interface would respond to different actions that they make. In a board game form of prototyping, the designers and developers can play the game alongside participants. This offers a richer embodied experience as demonstrated by the case studies. In both cases, we observed the use of representation and symbolization of the city (its space, place, and people) and used various metaphors to simulate interactions in a board game. In both case studies, effective use of board game play was used to symbolize the sociability and spatiality of the city as cultural metaphors of the everyday. They were

able to adopt board game as a tool for representing movement and context awareness, essential features of location-based games.

The paper prototypes were successful in incorporating freedom of movement and directionality but were limited when trying to represent the physical effort experienced by an individual who is traversing and discovering a real city. Although they did not sufficiently represent and manage the place attachments that we have with real places. They are a helpful tool for simulating movement in the city when time and resources do not support being there. They are a useful tool for co-designing elements of gamification and defining and redefining game rules. We also found that they are invaluable method for trialing content (form, style and quality). They are not without their limitations especially since the paper map is a simplified model of a complex reality—one that can only ever be partially modelled and the complexity. So when making the transition from the board game prototype one must be careful to consider these limitations and the different nature of the experience players have when they are out in the wilds of the city.

Acknowledgements The CrossCult (www.crosscult.eu) project Pilot 4 application was funded by the European Union's Horizon 2020 research and innovation programme under grant agreement no 693150.

Dr. Konstantinos Papangelis contributions have been supported through Xi'an Jiaotong-Liverpool University's RDF-15-02-17 programme.

References

- Ahlqvist O, Schlieder C (eds) (2018) *Geogames and Geoplay: game-based approaches to the analysis of geo-information*. In: *Advances in geographic information science*. Springer International Publishing
- Benford S, Crabtree A, Reeves S, Sheridan J, Dix A, Flintham M, Drozd A (2006) Designing for the opportunities and risks of staging digital experiences in public settings. *ACM Press*, p 427. <https://doi.org/10.1145/1124772.1124836>
- Chamberlain A, Bødker M, Papangelis K (2017) Mapping media and meaning: autoethnography as an approach to designing personal heritage soundscapes. *Proc Audio Most 2017 Augment Particip SoundMusic Exp*. <https://doi.org/10.1145/3123514.3123536>
- Chorley MJ, Whitaker RM, Allen SM (2015) Personality and location-based social networks. *Comput Hum Behav* 46:45–56. <https://doi.org/10.1016/j.chb.2014.12.038>
- de Souza e Silva A, Frith J (2012) *Mobile interfaces in public spaces: locational privacy, control, and urban sociability*. Routledge, New York
- Ehn P, Kyng M (1992) Cardboard computers: mocking-it-up or hands-on the future. In: *Design at work*. L. Erlbaum Associates Inc., pp 169–196
- Evans L, Saker M (2017) *Location-based social media: space, time and identity*. Springer International Publishing
- Flanagan M (2013) *Critical play: radical game design*. MIT Press, Cambridge
- Huizinga J (1955) *Homo ludens*. Beacon Press, Boston
- Jones CE, Liapis A, Lykourantzou I, Guido D (2017a) Board game prototyping to co-design a better location-based digital game. In: *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*. *ACM*, pp 1055–1064

- Jones CE, Liapis A, Lykourantzou I, Guido D (2017b) Board game prototyping to co-design a better location-based digital game. ACM Press, pp 1055–1064. <https://doi.org/10.1145/3027063.3053348>
- Jones CE, Theodosis S, Lykourantzou I (2019) The enthusiast, the interested, the sceptic & the cynic: understanding user experience & perceived value in location-based cultural heritage games through qualitative and sentiment analysis. ACM J Comput Cult Herit Spec Issue Eval Digit Cult Resour 11
- Li Y, Hong JI, Landay JA (2004) Topiary: a tool for prototyping location-enhanced applications. In: Proceedings of the 17th annual ACM symposium on user interface software and technology. ACM, pp 217–226
- Licoppe C, Inada Y (2009) Mediated proximity and its dangers in a location aware community: a case of ‘stalking.’ Digit Cityscapes Merging Digit Urban Play 100–128
- Marins DR, de OD Justo M, Xexeo GB, de AM Chaves B (2011) SmartRabbit: a mobile exergame using geolocation. In: 2011 Brazilian symposium on games and digital entertainment (SBGAMES). IEEE, pp 232–240
- Mateos MJ, Muñoz-Merino PJ, Kloos CD, Hernández-Leo D, Redondo-Martínez D (2016) Design and evaluation of a computer based game for education. In: Frontiers in education conference (FIE), 2016 IEEE. IEEE, pp 1–8
- Matyas S, Matyas C, Schlieder C, Kiefer P, Mitarai H, Kamata M (2008) Designing location-based mobile games with a purpose: collecting geospatial data with CityExplorer. ACM Press, p 244. <https://doi.org/10.1145/1501750.1501806>
- Nijholt A (ed) (2017) Playable cities: the city as a digital playground, gaming media and social effects. Springer Singapore
- Papangelis K, Metzger M, Sheng Y, Liang H-N, Chamberlain A, Cao T (2017a) Conquering the city: understanding perceptions of mobility and human territoriality in location-based mobile games. Proc ACM Interact Mob Wearable Ubiquitous Technol 1:1–24. <https://doi.org/10.1145/3130955>
- Papangelis K, Metzger M, Sheng Y, Liang H-N, Chamberlain A, Khan V-J (2017b) Get Off My Lawn!: starting to understand territoriality in location based mobile games. In: CHI EA '17 proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems. ACM Press, pp 1955–1961. <https://doi.org/10.1145/3027063.3053154>
- Saker M, Evans L (2016) Everyday life and locative play: an exploration of foursquare and playful engagements with space and place. Media Cult Soc 38:1169–1183
- Schlieder C, Kiefer P, Matyas S (2006) Geogames: designing location-based games from classic board games. IEEE Intell Syst 21:40–46. <https://doi.org/10.1109/MIS.2006.93>
- Snyder C (2003) Paper prototyping: the fast and easy way to design and refine user interfaces. The Morgan Kaufmann series in interactive technologies. Morgan Kaufmann Publisher, San Diego

Agent-Based Simulation for Indoor Manufacturing Environments—Evaluating the Effects of Spatialization



Stefan Kern and Johannes Scholz

Abstract The paper elaborates on an Agent-based Modeling approach for an indoor manufacturing environment—in particular, a semiconductor production plant. In order to maintain a flexible production “line”, there is no conveyor belt, and a mix of different products is present in the indoor environment. With the integration of Industry 4.0 or Smart Manufacturing principles, production assets may be transported by autonomous robots in the near future. The optimization of manufacturing processes is challenging and computationally hard. Thus, simulation methods are used to optimize manufacturing plants and the processes. In contemporary literature, the effects of the spatial dimension with respect to the simulation of manufacturing processes is neglected. In this paper, we evaluate on the effects the spatial dimension in an Agent-based Model for indoor manufacturing environments. The Agent-based Model developed in this paper is utilized to simulate a manufacturing environment with the help of an artificial indoor space and a set of test data. Four simulation scenarios—with varying levels of spatial data usable—have been tested using Repast Symphony framework. The results reveal that different levels of available spatial information have an influence on the simulation results of indoor manufacturing environments and processes. First, the distances moved by the worker agents can be significantly reduced and the unproductive movements of worker agents (without production assets) can be decreased.

Keywords Indoor geography · Smart manufacturing · Industry 4.0 · Agent-based modeling

S. Kern (✉) · J. Scholz
Institute of Geodesy, Graz University of Technology,
Steyrergasse 30, A-8010 Graz, Austria
e-mail: johannes.scholz@tugraz.at

S. Kern
e-mail: stefan.kern@outlook.at

© Springer Nature Switzerland AG 2020
P. Kyriakidis et al. (eds.), *Geospatial Technologies for Local and Regional
Development*, Lecture Notes in Geoinformation and Cartography,
https://doi.org/10.1007/978-3-030-14745-7_17

1 Introduction

In industrial manufacturing processes efficiency is one of the key factors to succeed in global markets that tend to be increasingly competitive. In recent years, the buzzwords Industry 4.0 or Smart Manufacturing are “spearheading” the fourth industrial revolution, that is based on digitization in manufacturing. The digitization approaches in any industry sector, includes Cyber-Physical Systems (CPS), Cloud Computing and methods from the Internet of Things (IoT). These approaches lead towards a digital integrated factory—i.e. Smart Factory Hermann et al. (2016); Plattform Industrie 4.0 (2018)—where machines and production assets and humans share data and communicate. Finally, smart manufacturing will end up in digitally connected manufacturing plants, where IT systems may significantly support the decisions of humans or enable robots to make a certain decision in an autonomous manner Henning et al. (2013). In this paper we focus on the manufacturing processes in the semiconductor industry that has some peculiarities mentioned in Scholz and Schabus (2014). We deal with an indoor environment, where manufacturing devices process production assets in a clean-room. The transport between the production steps is not fully automated or done with a conveyor belt. Thus, humans perform every transport process with the help of trolleys. Several hundred production steps are necessary to create a fully functioning computer chip. The production is characterized by a high variability of manufacturing machinery that is capable of performing several manufacturing processes, which leads to a highly complex system in combination with behavior of the involved humans. Thus, it is hardly possible to generate a deterministic spatial optimization model that helps to simulate and optimize the manufacturing processes. Simulation can significantly help in developing feasible and accurate schedule with less computational effort in comparison to some of the other techniques (e.g. Mazziotti and Horne Jr 1997), even for the job shop scheduling problems which are considered as NP-hard (e.g. Pinedo 1995).

In literature, there are several simulation approaches mentioned that simulate several production devices. Nevertheless, most of the papers neglect the geography of manufacturing environments—i.e. the spatial arrangement/layout of manufacturing plants. Scholz and Schabus (2015) and Scholz and Schabus (2014) elaborate on the role of Geography for Industry 4.0 purposes and conclude that the spatial dimension could be helpful for gathering new insights in manufacturing related data.

In this paper, we focus on a manufacturing simulation that utilizes Agent-based Modeling and Simulation (ABMS). We evaluate the effects of the integration geography on the simulation results. The research question of the paper is as follows: “Does the integration of the spatial dimension have an influence on the results of ABMS for indoor manufacturing environments?”. Hence, this paper does not elaborate on the computational complexity nor the increase of implementation complexity by incorporating the spatial dimension in the ABMS.

The remainder of the paper is organized as follows. Section 2 discusses the relevant literature, followed by the approach followed in this research project in Sect. 3. Section 4 describes the experiment conducted. The results and their discussion are to be found in Sects. 5 and 6, respectively.

2 Relevant Literature

This section is intended to discuss the relevant literature for this paper. Basically, we elaborate on literature covering modeling indoor space from a Geographic Information Science (GISc) point of view, ABMS, and Smart Manufacturing/Industry 4.0. The article is based on these three pillars and utilizes theory and methodologies from all three fields of expertise.

Modeling indoor space is an active research field in GISc as there are numerous recent publications in this field (Cheema 2018; Balasubramanian 2018; Diakit  and Zlatanova 2018; Knoth et al. 2018; Scholz and Schabus 2017; Schabus et al. 2017). Modeling indoor space got into focus due to the fact that a person resides 90% of their time inside a building (Klepeis et al. 2001; Jenkins et al. 1992). The first papers on modeling indoor space were published by Raubal and Worboys (1999) and Raubal (2001), where the latter elaborated on a wayfinding simulation with an agent-based approach. Yang and Worboys (2015) describe the generation of navigation graphs in indoor space, which can be utilized by e.g. a combined ontology—operations research approach to find the most suitable path (Scholz and Schabus 2017).

Agent-based Modeling (ABM) simulates actions of autonomous agents, and displays their behaviour and the dependencies between agents. The aim of ABMs is not to reach equilibrium, but to analyze how a system reacts and adapts to changed conditions (Macal and North 2010; Crooks and Heppenstall 2012). According to Mandl (2003) ABMs generally incorporate an environment with a spatial (geographical) dimension, as well as a number of objects in the environment, which are passive and can be recognized, created, destroyed, and changed by the agents. In ABMs the environment is recognized as a passive part, and agents serve as stimulant for environmental change processes (Batty et al. 2012). In an ABM, relations exist, which connect objects with agents allowing them to recognize, create, consume, transform, and manipulate objects. Van Berkel and Verburg (2012) mention an ABM in which the willingness of agents, their opportunities and choices are simulated. ABMs can simulate interactions between individuals, groups of individuals, but also between policies not being represented by individuals. This approach can be utilized in land cover and land use modeling (Br ndle et al. 2015). Grimm and Railsback (2012) and Grimm et al. (2010) elaborate on a protocol to describe and formalize AMBS in a standardized way. ABMS have been widely used for simulation purposes in industry (e.g. Monostori et al. 2006; Negmeldin and Eltawil 2015; Leit o 2009).

Literature in the field of Smart Manufacturing or Industry 4.0 deals with digitization in the industrial sector (Drath and Horch 2014; Albach et al. 2015; Hermann et al. 2016; Kagermann 2014). Hermann et al. (2016) in particular describe the basic

principles and the structure of an Industry 4.0 scenario. As IoT plays an important role in this field Mattern and Floerkemeier (2010) elaborated on the basics of IoT for industrial applications. Shrouf et al. (2014); Almada-Lobo (2016); Bi et al. (2014), Geng (2005), and Osswald et al. (2013) discuss the terms industry 4.0 and smart factory in detail.

ABM for manufacturing is laid out in detail in Paolucci and Sacile (2016). Negahban and Smith (2014) elaborate on the literature published regarding the simulation of a manufacturing system. Wang et al. (2016) highlight the importance of multi-ABMS for Industry 4.0. They advocate for a self-organized ABMS that is assisted by feedback and coordination originating from the accumulated data base. Additionally, they argue that an intelligent negotiation mechanism for agents to cooperate with each other can be based on ABMS that may help to optimize a smart manufacturing facility. The development of an hierarchical ABMS approach to simulate a manufacturing environment is presented in Mönch et al. (2003). Nevertheless, literature does not show that there has been a study that elaborates on the effects of the spatial dimension for manufacturing spaces. Contemporary papers show the benefits of simulation approaches, but do not integrate the spatial dimension in the simulation and/or decision making process.

3 Approach

The approach in this paper is based on the ABMS paradigm. Hence, the universe of discourse is modeled with the help of individual agents that move around an indoor manufacturing space. In general, we define an ABM—i.e. agents, environment and their interactions—and conduct a number of simulations, which are based on defined test scenarios. In detail, the operator-agents have different levels of information—e.g. full set of spatial information, missing spatial information of tools, missing spatial data on production assets, or no spatial information at all. We evaluate the effects of these simulation scenarios on the ABM. In particular, the results evaluate the effects on the walking distance of operators, time to complete the given production lot, and thus walking distance without products (non-productive).

The first step is to define the ABM, which is the basis for the experiments. The model contains of the basic elements of a manufacturing process, which are implemented utilizing concepts of Industry 4.0. The components of the model, as well as their functions, properties and relations are listed below and given in Fig. 1.

- **Products:** Products are processed in the factory. A predetermined sequence of production steps must be carried out on each product. Each step has to be carried out on one or more defined machines. The products cannot move autonomously, and thus must be transported by workers using trolleys. Each product is clearly identifiable and knows its own production status. The product knows its position in the factory and its work steps.

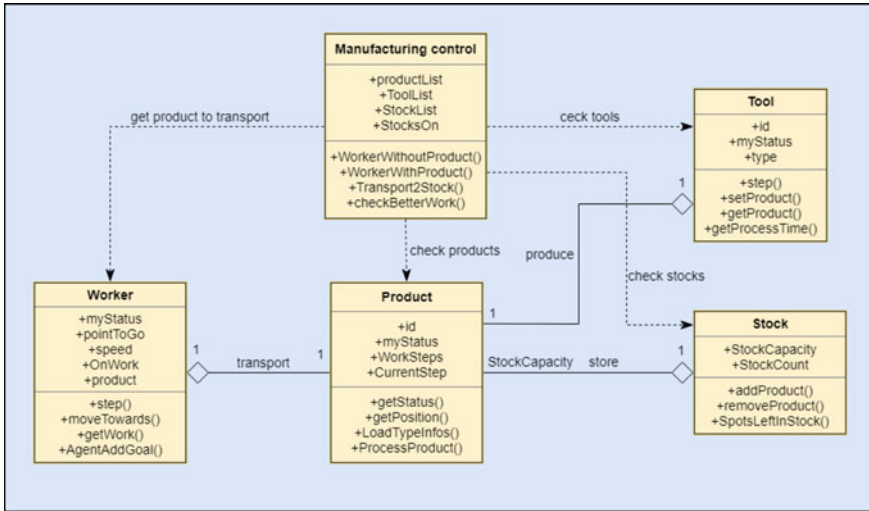


Fig. 1 Class diagram of the model with the most important functions and attributes of each agent class

- **Workers:** The workers or operators are the only mobile part of the model. They transport products through the factory. This means they must be equipped with the ability to navigate the factory on their own. Each worker can pick up and transport a single product, using a trolley. The workers are also able to search for transport orders on their own. Workers can communicate with the other components of the model and exchange information.
- **Machines/Tools:** The tools process the product and carry out a specific work step. This step is linked to the type of machine, so every machine of the same type can perform the same step. They have a fixed position in the plant. Machines know how much work they have ahead of them (i.e. products waiting to be processed), and share this information with other agents.
- **Stocks:** Intermediate storage serves as an intermediate step if there is no available machine for a certain product. They have a certain capacity, which can be queried.
- **Simulation environment:** A clean room in which the simulation is carried out. This indoor manufacturing facility represents the spatial dimension of the ABM. The structure of the factory with corridors and machines and stocks is designed similar to Scholz and Schabus (2014).

The model is implemented according to the concept of Industry 4.0 respectively a Smart Factory. This means that the basic elements of the model communicate with each other. Such factories have a decentralized manufacturing control, which means that there is not central instance that makes all decisions, but the entities in the manufacturing space can (up to a defined level) make their own decisions (Bi et al. 2014; Almada-Lobo 2016). The core element in a production process is the manufacturing control. The manufacturing control checks the processes in the

factory. The progress of the product is monitored while it moves through the various manufacturing steps in the factory. Decisions have to be made, such as:

- which production step of which product is done on which machine?
- when will it be released for onward transport?
- in which order should the products be produced?

Traditionally, such processes are hierarchically and centrally controlled Leitão (2009). With the number of required work steps as well as the number of interactions between departing components, the complexity of scheduling increases. And the system is becoming slower to adapt to new situations (Cantamessa 1997). Leitão (2009) provides a manufacturing control based on agents. This approach is based on the fact that all components of manufacturing are defined as agents. This means they work autonomously, are intelligent and work together. Thus, a product which is represented as an agent knows its past, knows which work step is next and can independently request further transport. These properties are also used in the description of an industry 4.0 scenario (Kagermann 2014; Albach et al. 2014). In the model proposed here, we developed a central administration of the manufacturing information. This administration does not plan or schedule the production process, but it acts as a mediator for requests from agents. Hence, the information administration provides e.g. the information to find the right manufacturing machine for a certain product.

3.1 Model Concept

Based on the definition of the components of the model, an overview of the planned simulation process follows. The processes of the model are depicted in Fig. 2. This figure shows the simulation process from a worker's point of view. The process is described in detail in the following list. In general, the ABM processes are applied to all worker agents and the processes are repeated until all products are finished.

- The process starts with workers without transportation tasks. As each worker shall transport products, the agent submits a transportation request to manufacturing control. The request contains the agent's position, which is required for the selection of the production asset. Using the worker's position, a so-called accessibility value is calculated for all products. This value is composed of the **distance** from the worker agent to the product and the remaining **production time** at the machine.

$$\text{accessibility value} = \frac{\text{distance}}{\text{speed}} + \text{remaining production time}$$

The distance is calculated using Dijkstra's shortest path algorithm (Dijkstra 1959), based on a network representation of the indoor manufacturing environment. Since the accessibility value is given in simulation ticks, the distance must still be divided by the speed of the workers. If the product is being processed on a machine,

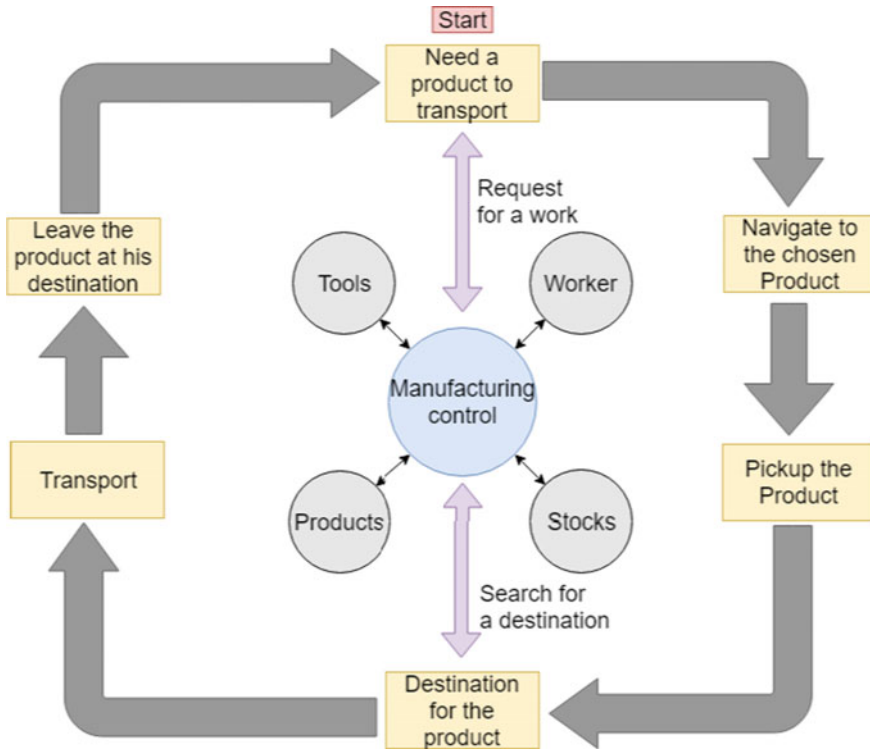


Fig. 2 Workflow of the agent-based production processes

remaining working time is added to accessibility value. The next product is selected via the accessibility value, the product with the lowest value is selected for the next transport task. This means that the production asset is selected that has the smallest the sum of travel and waiting time. The selected product is assigned to the workers for the transport. Once the product is assigned to a worker, it is no longer available to other workers until the transport process has been completed.

- After a worker agent has been assigned a product to transport, the agent navigates through the indoor manufacturing environment to the product.
- Once the worker agent has arrived at the product, he picks it up. The product can be located either at the starting point (i.e. where raw products enter the clean room), in an intermediate store or at a machine.
- Due to the fact that the product knows its next step, it is looking for a suitable machine for the next production step. A request is sent to the manufacturing control, which queries the production machines and forwards candidate machines to the production asset. Before doing that, it is checked whether all work steps have been completed on the product. If this is the case, the transport to the endpoint of the production takes place (i.e. the point where they leave the clean room). As soon as the product is at the end point, it is no longer relevant for the further simulation

and removed from the simulation. If the product is not ready yet, it is necessary to search for a suitable machine for the next step. Since there is normally more than one suitable machine, one particular needs to be selected. This is done similar to the product search via an **accessibility value**. The **accessibility value** is composed of the travel time to the machine and the remaining production time of the machine itself. The manufacturing machine having the minimum value is chosen. If no machine is available—happening when all machines are currently utilized by manufacturing processes—the transport will move the product to an intermediate store. If the stocks are full or deactivated in the simulation, the worker waits until a machine becomes available. If a machine has been selected as the destination for transport, it will be locked for other jobs during the transport process. It will not be released to other workers until the product has arrived at the machine. As a result, there are no redundant transports, and the scheduling process is simplified. The model offers the possibility for a worker to wait with a finished product in front of the machine while a production asset is currently being produced on a machine. We denote this situation as “exchange”, where the production asset in the machine is exchanged right after the manufacturing process is finished. To prevent a machine from being not available for the scheduling process for too long, the maximum waiting time can be controlled with the so-called *for exchange threshold*. This value is entered during the initialization of the simulation. And is given in simulation ticks. If the waiting time for a product exchange is greater than the specified threshold, the transport moves the product to an interim storage facility.

- The worker transports the product to the selected machine or an intermediate store.
- The product will leave at its destination. The worker is thus free again for new orders, and the cycle starts at the beginning.

The fact that the worker and not the product is the active part of the model has the following reasons. First, we model a semiconductor production facility Scholz and Schabus (2014) where humans perform the transportation tasks. Fully, autonomous transport is still not fully developed due to the high flexibility in the production layout and manufactured products. In addition, the production is not fulfilling the conveyor belt metaphor, which is done on purpose—to end up in a flexible manufacturing plant. Furthermore, if the product is transported using autonomous robots, the following question needs to be clarified: At what time will the transport request be made? A wrong choice may end up in serious delays or in a deadlock of the simulation system. This problem does not occur when a worker agent is looking for suitable products. Thus, the product being currently selected for transportation is available for the transportation process. This ensures that the worker is unladen for a short time only.

4 Experiment

The presented concept is implemented with Repast Symphony version 2.5¹ (North et al. 2013). For this purpose, the presented model is implemented and a test environment for simulation purposes is created. For the creation of the environment some considerations have to be made, which are listed as follows.

- **A layout of the production environment:** This layout contains the structure of the environment. On the one hand walkable areas such as paths, corridors, and open spaces (see Scholz and Schabus 2014), but also objects hindering movement, like walls or obstacles, are included. Another important issue for the production environment is the position of the machines and the intermediate storage, which was positioned similar to real semiconductor manufacturing spaces. To simulate the movement of the operators a network representation of the walkable space is generated, similar to Scholz and Schabus (2014).
- **Definition of the products:** For the products, a list of necessary manufacturing steps and transport speed for each product type are defined.
- **Definition of the machines:** The position and the type are determined by the environment layout (by placing similar machines in a corridor). In addition, the production time for a given manufacturing step on a machine is required.

Based on these considerations, the following test environment was generated. The created layout is shown in Fig. 3. In Fig. 3, the walls are black and impassable areas are gray. Tools are green, with four tools of the same type in each row. This results in 16 different machine types. The six squares in purple in the middle of the layouts are intermediate storage. The interim storage is therefore placed at the main corridor in front of the aisles with the tools. The point in magenta is the starting point, from here begins the production chain. The blue square on the right side is the end point, here the finished products are delivered and leave the clean room. The blue circle on the green square is the starting point of the works. At this point, the workers are placed at the initialization of the simulation. The generated map of the production environment has a size of 78×71 pixels. With a maximum of one pixel per step and a step length of one meter, the factory floor is 78×71 m.

The simulation of the described ABM above is done with several scenarios that are defined below. In addition we use the settings, defined in Table 1, throughout the simulation of the different scenarios. The scenarios are as follows:

- **With spatial information:** During this simulation run, workers, have spatial information about products and tools at their disposal. Also, the running times of the machines can be queried and included in the planning phase. Interim storage facilities are also available. This represents the developed optimal case of a smart factory simulation.
- **Without tool locations:** The workers, have no spatial information about the machines during this simulation run. However, the interim stocks can be used.

¹Repast Symphony 2.5: <https://repast.github.io/download.html>.

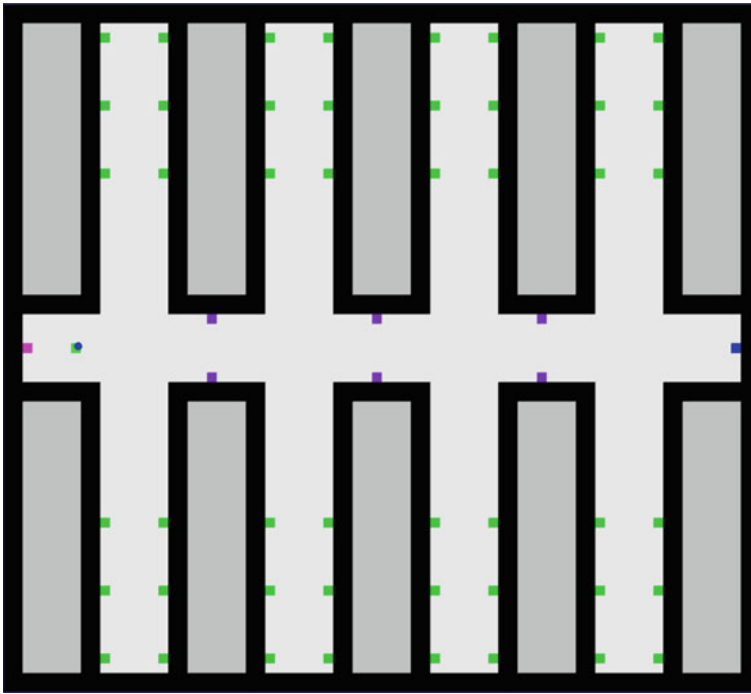


Fig. 3 The model of the indoor manufacturing space. This is the environment where agents move around during the simulation. The green dots are manufacturing machines, the purple ones denote stocks, the magenta one (on the left) represents the starting point of the production. The blue dot denotes the point where products leave the clean room

Table 1 Simulation settings of the ABMS for indoor manufacturing environments

Worker:	10
Products:	50
Stock capacity	15

The selected machine is the first to be free when searching for targets, the distance does not matter.

- **Without spatial information regarding products:** Here, the workers lack spatial information about the products. Here, all available products (i.e. currently not in a machine) are put in a list and the list items are randomized. The first element of the randomized list is selected for transport, regardless of the distance from worker to the product.
- **Without spatial information:** The worker/operator agent is not able to access spatial data for the machines and the products. Hence, the agent has to select and transport the first free product (from the randomized list of available products) to the first free machine.

Table 2 Comparison of the four simulation scenarios. Here the average runtime of each simulation scenario, the difference to the scenario “with spatial information”, and the average increase in % is given

	With spatial information	Without tool locations	Without product locations	Without spatial information
Runtime:	156030	165270	171104	178541
Difference:	0	9240	15074	22511
Increase %:		5.9%	9.7%	14.5%

Each scenario was simulated three times and the results were averaged. This is done in order to deal with the stochastic nature of the ABMs Crooks and Heppenstall (2012). In this model, as in most other ABMs the stochastic nature results from each agent’s ability to make autonomous decisions—which might be different at each single simulation run.

5 Results

The results of the experiments are shown in Table 2 and Fig. 4. As described earlier, the results are the average numbers of three test runs for each simulation scenario. The results depicted in Table 2 reveal the total runtime of the simulation and the difference between the runtime of the four test scenarios (in comparison to the scenario “with spatial information”). The results are sorted in ascending order of the runtime.

The influence of spatial information on the choice of the next product and transport targets becomes clear when looking at Fig. 5. This plot shows the distance traveled by the workers. It is to be recognized that without distance information to the destinations, the workers are on their way empty about half of their distance. Thus, it can be said that workers that utilize distance information for their decision making may increase the efficiency of the manufacturing processes. This is also supported by the results presented in Table 3. In the worst case 46.5% of the worker agents’ movements are done without carrying products (for the scenario “without spatial information”), whereas in the best case only 21.45% of the movement is done without products (scenario “with spatial information”). Additionally, the overall distance traveled is decreasing as well the more spatial information is used by the worker agents. The agents in the scenario “without spatial information” travel a distance of 65610 length units, whereas in the scenario “with spatial information” the agents move 41395 length units on the average. The overall smaller distance traveled by agents indicates that the spatial information may be relevant for indoor manufacturing simulation (Fig. 6).

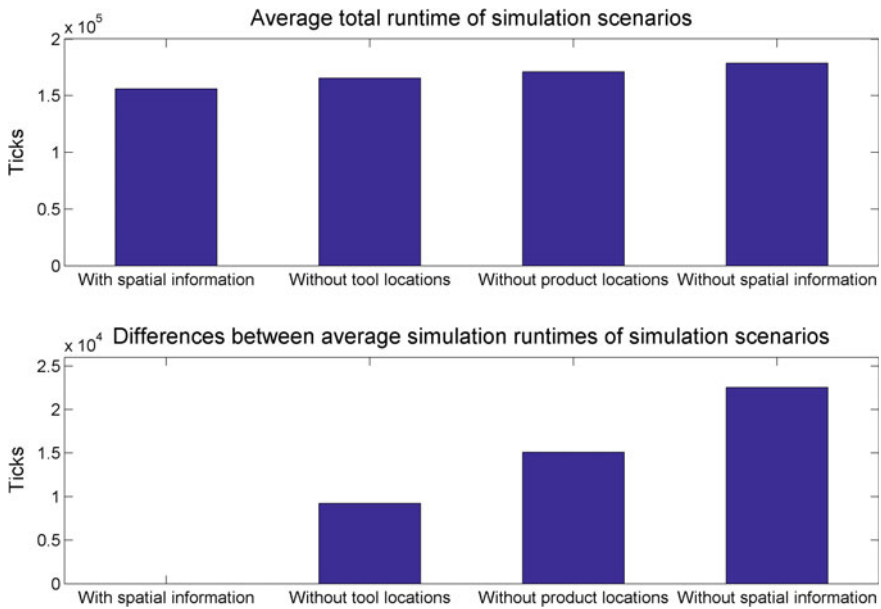


Fig. 4 Comparison of average simulation runtimes of the ABMS scenarios (above), and the differences of the runtimes of the ABMS runtimes for the scenario, with respect to the scenario “with spatial information” (below)

Table 3 Comparison of the four tested scenarios. The average distance covered by workers with and without products is considered. The data is given in general units of length

Worker status	With spatial information	Without tool locations	Without product locations	Without spatial information
with prod.	32515	35083	32350	35097
without prod.	8880	13518	25430	30513

6 Discussion and Conclusion

The paper deals with an ABM approach for an indoor manufacturing environment. In particular the ABM developed in this work models a complex indoor manufacturing plant, where workers transport the production assets from one manufacturing device to the next. This can be justified by that fact, that e.g. semiconductor manufacturing requires a flexible production “line” that cannot rely on a static conveyor belt. As a consequence, humans perform the transport tasks, up to now. With the integration of Industry 4.0 or smart manufacturing principles, production assets may be transported by autonomous robots. Nevertheless, the optimization of such complex manufacturing processes—especially true for the semiconductor industry—is challenging and computationally hard. Thus, simulation methods have been widely employed to opti-

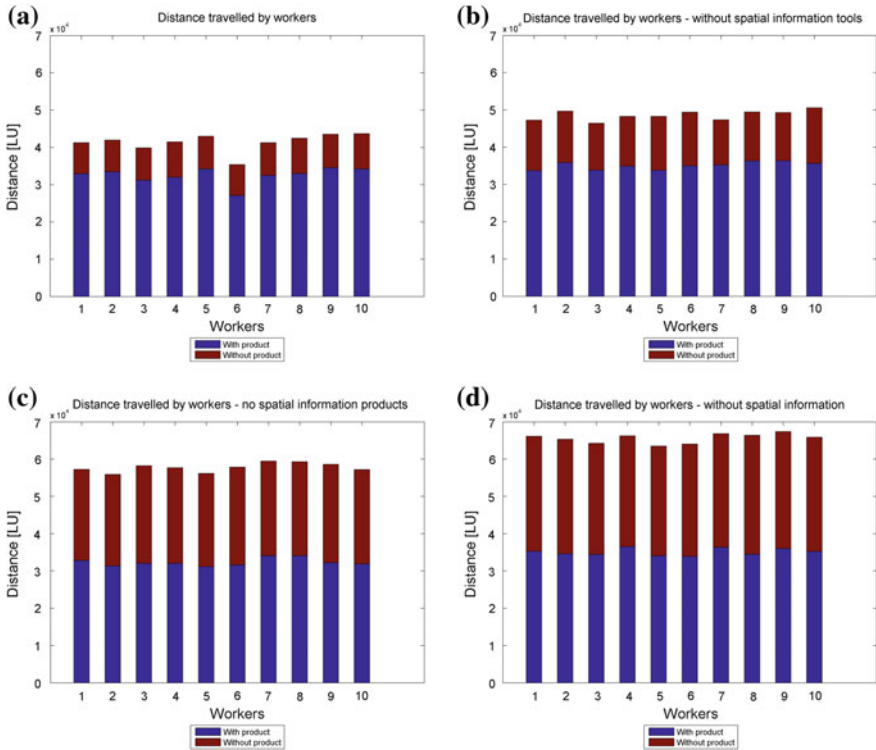


Fig. 5 Average distance moved by each worker agent, in distance units. With products is colored in blue and without products is depicted in red. Subfigure **a** shows the results for scenario “with spatial information”, subfigure **b** the scenario “without tool locations”, **c** the scenario “without product locations” and **d** the results of scenario “without spatial information”

mize manufacturing plants/processes. In particular ABMS is a micro-scale simulation that is capable of delivering results on a macro-scale.

In contemporary literature the effects of the spatial dimension of manufacturing processes is neglected—which is also true for a wide range of manufacturing simulations. In this paper, we evaluated the effects of incorporating the spatial dimension in an ABMS for indoor manufacturing environments, which is reflected in the research question: “Does the integration of the spatial dimension have an influence on the results of ABMS for indoor manufacturing environments?”. The ABMS developed in this paper simulates an manufacturing environment with the help of an artificial indoor space and a set of test data. Four simulation scenarios—with varying levels of spatial data usable for the worker agent, performing the transportation processes—have been tested using Repast Symphony framework. The results reveal that different levels of available spatial information have an influence on the simulation results of indoor manufacturing environments and processes. First, the distances moved by the worker agents can be significantly reduced—by approx. 34%, and the unproductive

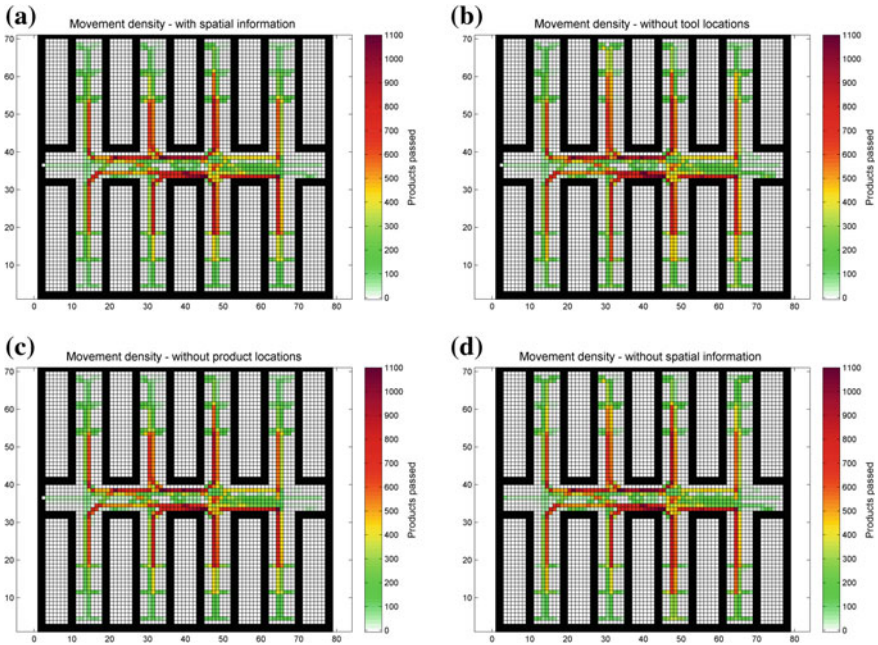


Fig. 6 Heatmap of the worker agents’ movements for the four simulation scenarios. Subfigure **a** shows the results for scenario “with spatial information”, subfigure **b** the scenario “without tool locations”, **c** the scenario “without product locations” and **d** the results of scenario “without spatial information”

movements (without production assets) can be decreased significantly. In addition, the computational complexity (i.e. run time) of the simulation lower, if the spatial dimension is included, which may be due to the fact that the manufacturing and transport processes are more efficiently aligned because of the spatial dimension.

References

- Albach H, Meffert H, Pinkwart A, Ralf R (2015) Management of permanent change. Springer, Berlin, pp 1–240
- Albach H, Meffert H, Pinkwart A, Reichwald R (2014) Management of permanent change—new challenges and opportunities for change management. Management of permanent change. Springer Fachmedien Wiesbaden, Wiesbaden, pp 3–21
- Almada-Lobo F (2016) The industry 4.0 revolution and the future of manufacturing execution systems (mes). *J Innov Manag* 3(4):16–21
- Balasubramanian V (2018) Geospatial infrastructure, applications and technologies. India case studies. In: *ingist: a queryable and configurable indoorgis toolkit*. Springer, Singapore, pp 93–105
- Batty M, Crooks AT, See LM, Heppenstall AJ (2012) Perspectives on agent-based models and geographical systems. *Agent-based models of geographical systems*. Springer, Berlin, pp 1–15

- Bi Z, Da Xu L, Wang C (2014) Internet of things for enterprise systems of modern manufacturing. *IEEE Trans Ind Inform* 10(2):1537–1546
- Brändle JM, Langendijk G, Peter S, Brunner SH, Huber R (2015) Sensitivity analysis of a land-use change model with and without agents to assess land abandonment and long-term re-forestation in a swiss mountain region. *Land* 4(2):475–512
- Cantamessa M (1997) Agent-balsed modeling and management of manufacturing systems. *Comput Ind* 34(97):173–186
- Cheema MA (2018) Indoor location-based services: challenges and opportunities. *SIGSPATIAL Spec* 10(2):10–17
- Crooks A, Heppenstall A (2012) Introduction to agent-based modelling. *Agent-based models of geographical systems*. Springer, Berlin, pp 85–96
- Diakité AA, Zlatanova S (2018) Spatial subdivision of complex indoor environments for 3d indoor navigation. *Int J Geogr Inf Sci* 32(2):213–235
- Dijkstra EW (1959) A note on two problems in connexion with graphs. *Numerische Mathematik* 1:269–271
- Drath R, Horch A (2014) Industrie 4.0: hit or hype? [industry forum]. *IEEE Ind Electron Mag* 8:56–58
- Geng H (2005) *Semiconductor manufacturing handbook*. McGraw-Hill, Inc, New York
- Grimm V, Berger U, DeAngelis DL, Polhill JG, Giske J, Railsback SE (2010) The ODD protocol: a review and first update. *Ecol Model* 221(23):2760–2768. <https://doi.org/10.1016/j.ecolmodel.2010.08.019>
- Grimm V, Railsback SS (2012) Designing, formulating, and communicating agent-based models. *Agent-based models of geographical systems*. Springer, Berlin, pp 361–377
- Henning K, Wolfgang W, Johannes H (2013) Recommendations for implementing the strategic initiative industrie 4.0. Final report of the Industrie 4.0 WG, 82
- Hermann M, Pentek T, Otto B (2016) Design principles for industrie 4.0 scenarios. In: *Proceedings of the annual Hawaii international conference on system sciences*, pp 3928–3937
- Jenkins PL, Phillips TJ, Mulberg EJ, Hui SP (1992) Activity patterns of californians: use of and proximity to indoor pollutant sources. *Atmos Environment Part A Gen Top* 26(12):2141–2148
- Kagermann H (2014) Change through digitization—value creation in the age of industry 4.0. *Management of permanent change*. Springer Fachmedien Wiesbaden, Wiesbaden, pp 23–45
- Klepeis NE, Nelson WC, Ott WR, Robinson JP, Tsang AM, Switzer P, Behar JV, Hern SC, Engelmann WH et al (2001) The national human activity pattern survey (nhaps): a resource for assessing exposure to environmental pollutants. *J Expo Anal Environ Epidemiol* 11(3):231–252
- Knöth L, Scholz J, Strobl J, Mittlböck M, Vockner B, Atzl C, Rajabifard A, Atazadeh B (2018) Cross-domain building models—a step towards interoperability. *ISPRS Int J Geo-Inf* 7(9):363
- Leitão P (2009) Agent-based distributed manufacturing control: a state-of-the-art survey. *Eng Appl Artif Intell* 22(7):979–991
- Macal CM, North MJ (2010) Tutorial on agent-based modelling and simulation. *J Simul* 4(3):151–162
- Mandl P (2003) Multi-agenten-simulation und raum - spielweise oder tragfähiger modellierungsansatz in der geographie. *Klagenfurter Geographische Schriften* 23:5–34
- Mattern F, Floerkemeier C (2010) From the internet of computers to the internet of things. From active data management to event-based systems and more, vol 6462. *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*. Springer, Berlin, pp 242–259
- Mazziotti BW, Horne RE Jr (1997) Creating a flexible, simulation-based finite scheduling tool. In: *Proceedings of the 29th conference on winter simulation*. IEEE Computer Society, pp 853–860
- Mönch L, Stehli M, Zimmermann J (2003) Fabmas: an agent-based system for production control of semiconductor manufacturing processes. In: *International conference on industrial applications of holoic and multi-agent systems*. Springer, pp 258–267
- Monostori L, Kumara S, Váncaza J (2006) Agent-based systems for manufacturing. *CIRP Ann Manuf Technol* 55(2):697–720 <http://www.sciencedirect.com/science/article/pii/S1660277306000053>

- Negahban A, Smith JS (2014) Simulation for manufacturing system design and operation: literature review and analysis. *J Manuf Syst* 33(2):241–261
- Negmeldin MA, Eltawil A (2015) Agent based modeling in factory planning and process control. In: 2015 IEEE international conference on industrial engineering and engineering management (IEEM), pp 1810–1814
- North MJ, Collier NT, Ozik J, Tataru ER, Macal CM, Bragen M, Sydelko P (2013) Complex adaptive systems modeling with repast simphony. *Complex Adapt Syst Model* 1(1):3
- Osswald S, Weiss A, Tscheligi M (2013) Designing wearable devices for the factory: rapid contextual experience prototyping. In: 2013 international conference on collaboration technologies and systems (CTS). IEEE, pp 517–521
- Paolucci M, Sacile R (2016) Agent-based manufacturing and control systems: new agile manufacturing solutions for achieving peak performance. CRC Press, Boca Raton
- Pinedo M (1995) Scheduling: theory, algorithms and applications. Prentice-Hall, Englewood Cliffs, NJ
- Plattform Industrie 4.0 (2018) <http://www.plattform-i40.de/I40/Navigation/DE/Home/home.html>
- Raubal M (2001) Ontology and epistemology for agent-based wayfinding simulation. *Int J Geogr Inf Sci* 15(7):653–665
- Raubal M, Worboys M (1999) A formal model of the process of wayfinding in built environments. *Spatial information theory. cognitive and computational foundations of geographic information science*. Springer, Berlin, pp 381–399
- Schabus S, Scholz J, Lampoltshammer TJ (2017) Mapping parallels between outdoor urban environments and indoor manufacturing environments. *ISPRS Int J Geo-Inf* 6(9):281
- Scholz J, Schabus S (2014) An indoor navigation ontology for production assets. In: Proceedings of the 8th international conference, GIScience 2014, Vienna, Austria, 24–26 September 2014, pp 204–220
- Scholz J, Schabus S, (2015) Geographic information science and technology as key approach to unveil the potential of industry 4.0. In: 2015 Proceedings of 12th International Conference on Informatics in Control, Automation and Robotic (ICINCO), vol.2, pp 463–470
- Scholz J, Schabus S (2017) Towards an affordance-based ad-hoc suitability network for indoor manufacturing transportation processes. *ISPRS Int J Geo-Inf* 6(9):280
- Shrouf F, Ordieres J, Miragliotta G (2014) Smart factories in industry 4.0: a review of the concept and of energy management approached in production based on the internet of things paradigm. In: 2015 IEEE international conference on industrial engineering and engineering management, pp 697–701
- Van Berkel DB, Verburg PH (2012) Combining exploratory scenarios and participatory backcasting: using an agent-based model in participatory policy design for a multi-functional landscape. *Landsc Ecol* 27(5):641–658
- Wang S, Wan J, Zhang D, Li D, Zhang C (2016) Towards smart factory for industry 4.0: a self-organized multi-agent system with big data based feedback and coordination. *Comput Netw* 101:158–168
- Yang L, Worboys M (2015) Generation of navigation graphs for indoor space. *Int J Geogr Inf Sci* 29(10):1737–1756

Part V
User and Workforce Dimensions
of Geospatial Technologies

Towards a Usability Scale for Participatory GIS



Andrea Ballatore, Will McClintock, Grace Goldberg and Werner Kuhn

Abstract Since its emergence in the 1990s, the area of Participatory GIS (PGIS) has generated numerous interactive mapping tools to support complex planning processes. The need to involve non-expert users makes the usability of these tools a crucial aspect that contributes to their success or failure. While many approaches and procedures have been proposed to assess usability in general, to date there is no standardized way to measure the overall usability of a PGIS. For this purpose, we introduce the Participatory GIS Usability Scale (PGUS), a questionnaire to evaluate the usability of a PGIS along five dimensions (user interface, spatial interface, learnability, effectiveness, and communication). The questionnaire was developed in collaboration with the user community of SeaSketch, a web-based platform for marine spatial planning. PGUS quantifies the subjective perception of usability on a scale between 0 and 100, facilitating the rapid evaluation and comparison between PGIS. As a case study, the PGUS was used to collect feedback from 175 SeaSketch users, highlighting the usability strengths and weaknesses of the platform.

Keywords Participatory GIS · PGIS · Participatory GIS Usability Scale (PGUS) · Usability evaluation · User experience · Web mapping

A. Ballatore (✉)

Department of Geography, Birkbeck, University of London, London, UK

e-mail: a.ballatore@bkk.ac.uk

W. McClintock · G. Goldberg

Marine Science Institute, University of California, Santa Barbara, Santa Barbara, CA, USA

e-mail: mcclintock@msi.ucsb.edu

G. Goldberg

e-mail: grace.goldberg@ucsb.edu

W. Kuhn

Department of Geography, Center for Spatial Studies, University of California,

Santa Barbara, Santa Barbara, CA, USA

e-mail: kuhn@geog.ucsb.edu

© Springer Nature Switzerland AG 2020

P. Kyriakidis et al. (eds.), *Geospatial Technologies for Local and Regional*

Development, Lecture Notes in Geoinformation and Cartography,

https://doi.org/10.1007/978-3-030-14745-7_18

1 Introduction

The term Participatory GIS (PGIS) refers to the usage of digital mapping tools to increase participation in planning processes and political negotiation, both in public and private contexts (Schlossberg and Shuford 2005; Sieber 2006; Rinner et al. 2008). Over the past 20 years, groups of planners, politicians, activists, and citizens have engaged in forms of computer-aided (or mediated) activities revolving around complex spatial problems that involve conflicting views and claims (Kwaku Kyem 2004). A typical PGIS is able to store, visualize, analyze, and annotate spatial objects in a collaborative manner, often with a strong cartographic component. Such systems are designed to enable multiple stakeholders to create and edit data and interact with other participants (Dunn 2007). PGIS application domains include land use zoning (Brown et al. 2018), urban planning (Maquil et al. 2018), indigenous land rights (Brown and Kytä 2018), assessment of environmental impact (Evans et al. 2004), and marine spatial planning (Mare Nostrum 2016).

While PGIS projects in the 1990s ran on unashamedly hard-to-use desktop GIS manned by experts, the following 15 years witnessed a boom of increasingly cheap and portable devices, coupled with the rise of ubiquitous web maps (Haklay et al. 2008; Smith 2016). Many PGIS project assemble ad-hoc systems using existing Web-mapping platforms,¹ while dedicated platforms include Commonplace.is and SeaSketch.² In parallel, the diffusion of smartphones since 2007 has created a novel category of computing platforms, characterized by smaller screens, sensors, touch interfaces, and well-integrated native apps.³ These rapid technological changes generally raised the expectations of users that interact with user-friendly, reliable, consistent, responsive, and aesthetically pleasing systems. As a result, the issue of usability is not an afterthought, but a central concern in an increasingly competitive software ecosystem.

Originating in the study of ergonomics and human-computer interaction, the term usability refers to how easy a user interface is to use, with respect to learnability, efficiency, memorability, errors, satisfaction, and other dimensions (Nielsen 1999, 2012a). Developers across domains have pledged to adopt user-centered design, emphasizing the needs of the user as a driving factor—or at least paying lip service to this idea. Usability engineering, user experience (UX), and their cognate fields are enjoying considerable popularity, and diverse methods and techniques are blossoming in a growing marketplace (Hassenzahl and Tractinsky 2006; Garrett 2010).

The context of PGIS has peculiarities that clearly distinguish it from other applications and domains. Unlike large commercial websites, PGIS tend to have a limited number of users, usually on the order of hundreds or thousands at most. The development of PGIS tools tends to be performed by small teams on tight budgets, in the framework of academic, non-profit, or governmental projects. As disadvantaged

¹<http://www.ppgis.net/resources/geoweb-applications>—All URLs were accessed in November 2018 and can be found on the Wayback Machine at <https://archive.org/web>.

²<https://www.commonplace.is>, <https://www.seasketch.org>.

³See for example ESRI's apps at <http://www.esri.com/software/apps>.

communities are often the focus of PGIS initiatives, users' digital literacy might vary dramatically. It is also important to note that, in projects using PGIS, social, political, and cultural obstacles can be harder to overcome than tools' usability (Brown et al. 2017). Despite the general interest in usability engineering in GIScience (Slocum et al. 2001; Raubal 2009), our knowledge of usability for PGIS remain fragmented and, because of technical changes, rapidly obsolescent (Haklay 2010).

For this reason, we argue that specific usability guidelines and approaches should be developed for PGIS, coordinating the sparse efforts in the field. Considering the typical technical and institutional context of PGIS projects, we aim at developing a PGIS usability evaluation framework, with a set of clear guidelines, practices, and tools that will help managers and developers. As the first step in this direction, we focus on the issue of the measurement of the perception of overall usability of a PGIS.

Quantifying the subjective perception of usability has long been identified as a key aspect of usability engineering (Brooke 2013). If users find a system not usable, these impressions will impact negatively on the system's adoption, regardless of other, more objectively measurable, aspects. In this context, the System Usability Scale (SUS), designed by a usability engineer in the 1980s, emerged as simple and useful scheme to measure the overall level of usability of a software system (Brooke 1996). Based on ten questions, the SUS captures in a repeatable and consistent way how usable a system is perceived to be, and enables rapid comparisons with previous or competing systems. The questionnaire has stood up to scrutiny of statisticians and usability researchers for 30 years, without revealing fundamental flaws (Sauro 2011).

Considering the limited resources and constraints of PGIS projects, the emphasis should be on inexpensive approaches, providing practical support and guidance to developers, without the need for complex protocols and laboratory experiments. This family of *lean* or *discount* techniques has been promoted by several leading usability researchers, arguing that frequent, informal evaluations are to be preferred to rare, expensive, and formal ones (Nielsen 2009; Krug 2014). Following this lean approach to usability, we tackled the problem of measuring the user perception of usability in PGIS.

Starting from the SUS, we designed a set of questions that capture central aspects of PGIS with simple, unambiguous language. This process resulted in the Participatory GIS Usability Scale (PGUS), a freely-available and general questionnaire that can be deployed by any PGIS practitioner. This questionnaire is available online under a Creative Commons license to be used, updated, refined, and extended by PGIS practitioners.⁴ Indeed, this method is not meant to replace existing usability assessment approaches, but rather to provide a complementary low-cost tool.

The PGUS was developed and evaluated on SeaSketch, a PGIS for marine spatial planning created at the University of California, Santa Barbara.⁵ SeaSketch is an entirely web-based platform for collaborative spatial planning and analytics, aimed

⁴<http://github.com/andrea-ballatore/pgis-usability>.

⁵<http://www.seasketch.org>.

at facilitating the creation, evaluation, and sharing of spatial sketches, i.e., map-based scenarios that intend to capture and communicate the user's viewpoint on planning problems. The system is currently deployed in several planning projects, and has, relatively to the field of PGIS, a large user base of about 3,200 people. The PGUS was developed iteratively with the collaboration and input of SeaSketch users. A set of 175 responses was obtained from the community, generally positive feedback about the tool's usability, while also identifying several areas of improvement that informed the team's development work.

In the remainder of this paper, we review the state of the art in the usability of PGIS (Sect. 2). SeaSketch provided a suitable case study (Sect. 3). Subsequently, we lay the groundwork for a lean usability method for PGIS (Sect. 4). Section 5 outlines the Participatory GIS Usability Scale (PGUS), while Sect. 6 describes the results of the PGUS applied to SeaSketch. Finally, we draw conclusions from this experience and we indicate directions for future work (Sect. 7).

2 Usability, Web Design, and PGIS

The need for better usability in PGIS has been widely acknowledged (Haklay and Tobón 2003; Aditya 2010; Skarlatidou et al. 2013). The cost of poor usability in PGIS projects is high, as participants can quickly come to believe that the new technology might not support the decision-making process or even hamper it (Haklay 2010). Project funders might regret the investment and reduce future support for similar initiatives. Engineering high usability for a PGIS is therefore of crucial importance to avoid such unfortunate situations.

2.1 *Measuring Usability*

Producing reliable observations on the complex interactions between people and machines presents many challenges. Usability is not directly observable, and is articulated in several interrelated dimensions, such as efficiency (the speed of execution of tasks), effectiveness (the quality of the solutions produced through the system), satisfaction (the emotions experienced when using a system), and learnability (the ease to remember procedures on the system) (Nielsen 1999). The conceptual organization of these dimensions and their priority varies across research communities and individual authors. Usability engineers often rely on qualitative methods to identify specific issues, and on quantitative approaches to detect broad trends that would not be observable qualitatively (Nielsen 2004).

As Nielsen (2012b) pointed out, some aspects of usability are measurable by considering objective outcomes of user behavior, while others depend on the subjective and affective relationship of the user with the system. For instance, the efficiency of a system in supporting a task can be measured in terms of success rate and elapsed

time. By contrast, the measurement of user satisfaction must involve subjective judgments. Unsurprisingly, users tend to be more satisfied with better-performing systems, but the correlation between objective performance and subjective satisfaction is not absolute, requiring distinct measurement strategies (Frøkjær et al. 2000).

The need to study systematically these objective and subjective dimensions prompted the design of several metrics and questionnaires (Hornbæk 2006). Notably, usability experts can make users perform tasks, counting their successes and failures in a binary way, studying their recurring errors, or looking the quality and completeness of their solutions to non-binary problems. Usability metrics can capture the precision of manipulation of interface elements (i.e., the ratio between intentional and unintentional actions on the interface). The learnability of a system can be quantified through measures of recall (the ability of the users to remember how to perform tasks or to recall information seen on the system's interface). Tasks can be designed to measure the user's cognitive load and efforts when communicating system-based procedures to other users.

To measure subjective dimensions of usability, questionnaires are fundamental tools. In particular, the standardization of questionnaires has been identified as an important way to measure success over time and compare systems (Hornbæk 2006). For general usability, several questionnaires have emerged as de facto standards, and have been applied to large numbers of systems and users. Notably, Lewis designed questionnaires for IBM in the 1990s, which have been widely used (Lewis 1995). Brooke's System Usability Scale (SUS) (Brooke 1996, 2013; Sauro 2011) is one of the most successful usability questionnaires, and provided the template for our work that aims at providing a standardized questionnaire for PGIS. The SUS was selected over competing questionnaires because of its simplicity and high statistical reliability.

2.2 Usability and Web Design

As Norman (2013) claims, many usability principles remain valid over time and in diverse application areas. However, as web browsers become increasingly powerful and influenced by the adoption of smartphones, the web is changing rapidly as an interactive medium. Standards like HTML5, CSS3, and JavaScript frameworks enable the design of full-fledged web-based applications, often with strong geospatial components. Therefore, it is reasonable to assume that PGIS will increasingly be web-based, and these trends should be considered with particular attention to design usable PGIS.

In recent years, web technologies have greatly improved, and great strides in the usability of websites have been made (Nielsen and Budiu 2013). As happened previously with desktop software development, web development increasingly relies on frameworks that support the design of standardized, consistent interfaces. Furthermore, in recent years, large actors in the US have developed their own usability philosophies, such as Apple's iOS Human Interface Guidelines, Google's Material

Design, and the US Government's Usability.gov. These guidelines have wide-ranging effects, well beyond the respective corporate platforms and websites.

Seemingly commonsensical principles of web design have been challenged by usability and UX research. The lighthearted website *UX Myths* provides a useful summary of such received ideas that appear to have no empirical basis.⁶ For instance, the *three-click rule* states that a user of a website should be able to find any information with no more than three mouse clicks, while this shows no correlation with performance or user satisfaction in real systems. Moreover, trends such as the so-called responsive design suggest to re-think interfaces as fluidly adaptable to any device, ranging from high-resolution large screens to typical smartphones, and even to smart watches (Gardner 2011). Maintaining high usability across radically different devices confronts PGIS developers and designers with new, cross-media challenges.

2.3 Usability of PGIS

PGIS is not a well-defined area, and its boundaries are often difficult to discern. Preferring the term Public and Participatory GIS (PPGIS), geographers and urban planners aim at understanding the interplay of the social, institutional, political, and technological dimensions in PGIS projects (Sieber 2006; Balram and Dragičević 2005; Forrester and Cinderby 2013).

While an account of these debates lies outside of the scope of this article, it is worth noting that PGIS is studied in its social and political effects, questioning whether it actually supports marginalized groups and social and environmental justice in real-world contexts (Brown 2012). Indeed, higher usability cannot overcome deeper constraints and barriers by itself, but non-usable PGIS have a slimmer chance to succeed.

Among the dozens of books published about usability over the past 20 years (e.g. Norman 2013; Krug 2014), only a textbook edited by Haklay (2010) focuses on the specific context of GIS, providing a valuable overview of the field, guidelines and case studies. In parallel, the usability of web maps, central to modern PGIS, has received some attention (Nivala et al. 2008; Roth 2015). Unfortunately, the highly specific context of PGIS studies and the rapid changes in web and mobile technologies limit the generalizability and currency of these findings.

The rare research in PGIS usability highlights that participants' previous web and GIS experience tends to affect the perceived usability of a system, confirming the challenges of usability design for participants with low levels of digital literacy (Stevens et al. 2014). Occasionally, in highly collaborative contexts, "old" media such as translucent maps appear more usable than "new" mobile interactive maps (Aditya 2010). In tense political contexts, the issue of trust has also received attention (Skarlatidou et al. 2013), and "serious games" have been proposed to add a playful component to PGIS, increasing public engagement (Poplin 2012).

⁶<http://uxmyths.com>.

Desktop-based systems such as ArcGIS are influenced by conventions and choices of their native platforms (i.e. Microsoft Windows), and the same principle applies to current GI web and smartphone apps, which are increasingly influenced by Google and Apple guidelines and constraints. Notably, Google Maps has become the de facto standard for reference maps in web applications, determining how hundreds of millions of users experience panning, zooming, and search on digital maps. It is still unclear how PGIS design practices should respond to these developments and, as a result, PGIS developers have to make several technical choices.

Hence, we argue that a new wave of usability research is needed to adapt to these recent changes and provide developers with lean, adaptive techniques to maximize the usability of PGIS, within its usually tight institutional and financial constraints. This article initiates this process by tackling the lack of standardized questionnaires for PGIS usability. The next section outlines SeaSketch, a PGIS that provides the real-world context to develop our usability framework and the PGUS.

3 SeaSketch, a Web Platform for Marine Spatial Planning

SeaSketch is a web-based software service mapping platform that facilitates the planning of ocean space, and is owned by the McClintock Lab at University of California, Santa Barbara.⁷ The SeaSketch platform was conceived as a successor to MarineMap, an application used for the collaborative design of marine protected areas in California (Gleason et al. 2010; Merrifield et al. 2013; Goldberg et al. 2016). The platform aims at supporting science-based, stakeholder-driven marine spatial planning in collaborative planning processes.

Marine Spatial Planning

Marine spatial planning, also referred to as ocean zoning, has as a main goal the coordination of stakeholders in the regulation of the usage of limited coastal resources, combining often divergent economic, social, and environmental objectives (Ehler and Douvère 2009). For example, marine spatial planning can be used in a coastal region to help fishing companies, local government, and environmentalists generate and analyze different planning options, defining sustainable fishing areas, while protecting endangered species in marine sanctuaries. Marine spatial planning can decrease conflict between stakeholders, improve regulatory efficiency, engage affected communities, and preserve fragile ecosystems, particularly in complex geopolitical contexts (Mare Nostrum 2016). To achieve these ambitious goals, marine spatial planning benefits from transparency and inclusiveness, engaging relevant stakeholders throughout the process.

SeaSketch Projects

To date, SeaSketch has been implemented and deployed in over twenty, large-scale planning initiatives including the Blue Halo Initiative in the Caribbean, the Southeast

⁷<http://www.seasketch.org>.

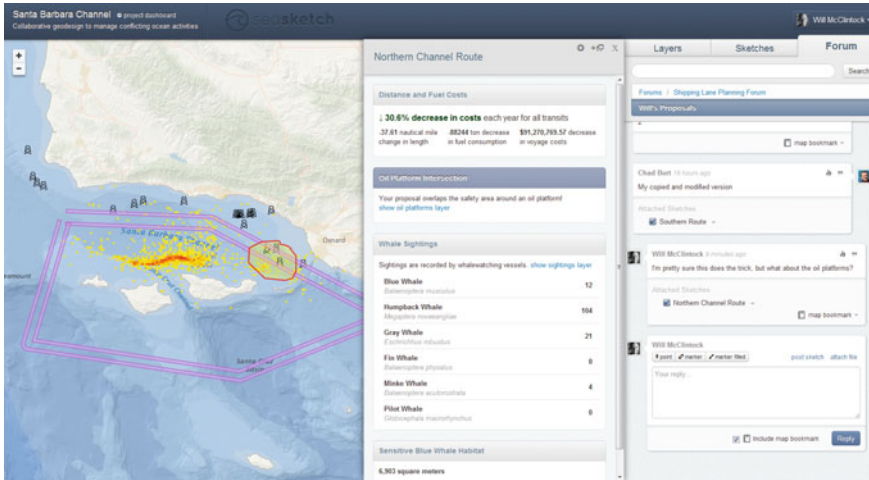


Fig. 1 The SeaSketch interface for collaborative marine spatial planning in the Santa Barbara Channel

Marine Protection Forum and SeaChange in New Zealand, the Marine Planning Partnership of the North Pacific Coast in British Columbia, the Marine Reserve Zoning process of Galapagos National Park, and the Safe Passages initiative in the Santa Barbara Channel of California.⁸ In each case, SeaSketch is used to facilitate face-to-face meetings, in which planners engage stakeholders or members of the general public in the design and evaluation of spatial plans. In most cases, the tool is also used by non-technical stakeholders at home or in community meetings without the assistance of planners.

The Tool

SeaSketch, launched in 2012, was designed for marine spatial planning to collect, evaluate, and analyze data, assessing whether scenarios met planning goals and objectives. The development team has collected feedback from users through the User Voice service,⁹ acting upon comments and feature requests, as part of an Agile development process. Figure 1 shows the interface of a SeaSketch-based system tailored to support collaborative marine spatial planning in the Santa Barbara Channel in California.

Users are invited to view relevant data and information, take surveys, and sketch out their plan ideas. These prospective scenarios are automatically evaluated based on science and policy guidelines and may be shared with other users in map-based discussion forums. These iterative design and analysis, often called geodesign (Goodchild 2010), encourage users to collaboratively build scenarios with broad stakeholder support.

⁸For a complete and updated list of projects, see <http://www.seasketch.org/projects>.

⁹<http://uservoice.com>.

In terms of interaction with spatial information, typical operations for planners and stakeholders include answering surveys to contribute to spatial datasets about human activities and resources in and around the ocean, viewing and querying map layers, sketching plans (e.g., marine protected areas, shipping lanes, tourism zones), and sharing and discussing plans in map-based forums. Project administrators, in addition, have control over the map extent, map layers, users and groups with access permissions, discussion forums, surveys, and they can analyze user activity in a project dashboard.

Usability of SeaSketch

While SeaSketch has been overall successful in its many deployments, the user feedback collected through the Agile development suggests that several aspects of its usability might be improved. For example, some users seem to be confused about how to sketch and analyze geometries, despite the contextual help intended to guide them through the process. Furthermore, user forums have seen limited usage and appear underutilized, probably for lack of clear feedback about privacy settings. Before addressing these specific issues, we realized that there was no simple and yet structured way to probe the perceived usability, and SeaSketch and its user base provide the ideal ground to advance the general knowledge about usability for PGIS. Lean usability, we argue in the next section, provides a suitable framework for this goal.

4 A Lean Usability Framework for PGIS

In a fast-changing technical landscape, usability for PGIS needs to be highly adaptive and flexible. In this section, we outline a lean usability framework for PGIS, based on our experience with the SeaSketch platform and user community. By lean usability, we mean usability engineering and evaluation techniques that favor informal, continuous, multi-modal, small-scale, iterative approaches to evaluating PGIS tools. Our lean usability framework has the following objectives:

- Evaluate the overall usability of the system to a given group of users.
- Help developers identify usability issues in current products and gain actionable insights from their users' behavior.
- Tighten the feedback loop between users and developers, tapping multi-modal computer-mediated communication.
- Embed continuous usability evaluation into the platform itself, not only in the form of ad-hoc interventions.
- Support design choices with evidence collected from users in natural settings.

Espousing Nielsen's viewpoint (Nielsen 2004), we argue that typical PGIS projects do not have enough resources to generate reliable findings, such as in A/B studies, where alternative versions of an interface are compared on a randomized sample of users. Lab-based evaluations are complex and expensive, and fail to capture

the diverse user bases and natural contexts of application of PGIS. Further principles include that observing what users do reveals more than asking users what they want (Krug 2014), and that implicit feedback analysis offers opportunities too (Ballatore and Bertolotto 2011). Based on the usability testing classification by Rohrer (2008), we propose a combination of the following methods for a typical PGIS.

Heuristic usability evaluations. A usability expert, external to the developers' team, reviews the products informally based on general usability principles such as consistency and standards, error prevention, and recognition rather than recall. Such evaluations can be seen as a premise to the other techniques.

User feedback buttons. Feedback buttons and links are placed in the products to allow users to express opinions, and can be closed- or open-ended. Feedback forms must be contextual, allowing users to comment on specific parts of the interface or tasks, rather than on the product in general. The entire feedback procedure must be extremely fast (usually shorter than 30 s), and should encourage informality.

Intercept surveys. Short and contextual web questionnaires can be triggered within the product, asking users focused questions, without making them leave the application. Compared to email surveys, intercept surveys capture feedback from users while they are performing tasks. To limit the impact on user attention, intercept surveys must be triggered rarely and preferably not during sensitive tasks. If triggered at the beginning or end of sessions, such surveys can be used to collect the intentions of users, illuminating their behavior.

Email surveys. Traditional email surveys can be used as a complement to intercept surveys to reach different users, particularly those who use the products infrequently. While the response rate is expected to be lower than for intercept surveys, email surveys can be longer and more detailed. The PGUS was disseminated as an email survey.

Implicit feedback. Modern web analytics packages can provide insights about user behavior, quietly collecting information about users' geographic location and devices, page views, clicks, taps, and even mouse movements. Implicit feedback analysis can reveal recurring usage patterns, providing actionable evidence to developers, for example suggesting the removal of an unused button or a shortcut to reach a frequently used panel. Concerns for user privacy are obviously paramount in PGIS and need to be addressed explicitly with an opt-in model.

To implement our first objective, that is the evaluation of the overall usability of the system, we designed a usability scale for PGIS, described in the next section.

5 Design of the Participatory GIS Usability Scale (PGUS)

As a first step towards the development of a usability framework for PGIS, we started by quantifying the perceived usability of a system on a numeric scale. Measuring the overall usability of a system is an obvious starting point to then proceed to other diagnostic techniques, and facilitates direct comparison between similar systems.

System Usability Scale (SUS)

In usability research, the System Usability Scale (SUS), originally conceived in 1986, has established itself as a standard tool for this kind of measurement (Brooke 2013). In SUS, users are asked to score ten items such as “I think that I would like to use this system frequently” and “I found the system unnecessarily complex” on a 5-point Likert scale, ranging from “strongly agree” to “strongly disagree.” These scores are then aggregated and converted into a final score from 0 (unusable) to 100 (extremely usable). SUS has been used in research and industry for 30 years, and is considered extremely general and reliable (Bangor et al. 2008; Sauro 2011).

Design Process

PGUS was designed in collaboration with members of the SeaSketch user community. The design of SUS involved the evaluation of two systems evaluated by participants, who filled in an initial questionnaire of 50 questions. From these questions, ten were selected based on inter-correlations. In the design process of PGUS, we did not have multiple systems to compare and, for this reason, we adopted a different approach. Starting from an initial set of 40 questions, we ask a sample of 16 users to provide feedback on the questionnaire overall, and specifically on each question. The user sample was designed to include expert administrators, intermediate users, and novices. Questions flagged as unclear, ambiguous or irrelevant were either rephrased, edited, or removed. After two iterations, the questionnaire was streamlined to 25 questions, marked as clear from all sample users. To maximize the response rate of the PGUS, we kept the maximum time of completion at about 5 min, deemed acceptable by the sample users (see Sect. 6).

Questions

The resulting questionnaire contains 25 questions, formulated in simple English, explicitly avoiding GIS jargon (see Table 1). It is worth noting that, although the SUS questions include positive and negative phrasing, we only included positive questions. Not mixing positive and negative phrasing reduces the possibility of misinterpretation and user error (Brooke 2013), making the score calculation simpler. However, it is important to note that this approach might reduce the participant’s attention devoted to each question.

The questions aim at capturing aspects that are common across PGIS in different domains. The questions are grouped in five themes, which represent complementary dimensions of the usability of a PGIS. These themes include (i) *user interface* (the visible part of the system), (ii) *spatial interface* (the interaction with spatial data), (iii) *learnability* (how easy it is to learn how to use the system), (iv) *effectiveness* (how the system supports the user goals), and (v) *communication* (how the system supports communication with other users and stakeholders). The names of the themes should not be displayed to the respondents taking the questionnaire, and are only for internal use. To reduce ordering bias, we recommend randomizing the order of the questions.

Table 1 Participatory GIS Usability Scale (PGUS)—v1

(A) <i>User interface</i>
1. The terms used in the system are clear.
2. It is easy to move through different parts of the system.
3. The error messages are easy to understand.
4. The delay between operations is acceptable.
5. Returning to the homepage is easy.
(B) <i>Spatial interface</i>
1. It is easy to move to a new location on the map.
2. It is easy to zoom in and out on the map.
3. I can create new content easily.
4. I can easily access information about what is displayed in the map.
5. The visual edits on the map take effect immediately.
(C) <i>Learnability</i>
1. I am confident using the system.
2. It is easy to remember how to perform tasks.
3. Discovering new features by trial and error is easy.
4. I find the help resources useful.
5. Mistakes can be easily undone.
(D) <i>Effectiveness</i>
1. The system gives me the tools to reach my goals.
2. The system is reliable.
3. I can complete tasks that would be impossible without the system.
4. The system increases my participation in the project.
5. I would recommend this system to others.
(E) <i>Communication</i>
1. The system helps me communicate my ideas to other participants.
2. I always understand what the system is showing.
3. The maps are easy to understand.
4. I can express my opinion about other participants' ideas.
5. When I have a problem, somebody can help me.

Scoring

Following SUS, the answer to a question is expressed on 5-point Likert scale, with each point corresponding to a score between 0 and 4: “Strongly disagree” (0), “Disagree” (1), “Neither agree nor disagree” (2), “Agree” (3), and “Strongly agree” (4). The final PGUS score, in analogy with SUS, is calculated by summing all the ratings, and ranges between 0 (unusable) and 100 (extremely usable). By keeping the scoring as simple as possible, PGUS avoids one of the common criticisms to SUS, reducing the possibility of error. The practitioner should bear in mind that the score does not represent a percentage of usability, and score comparisons across systems should be done by treating the data as ranked.

Resources and License

The current version (v1) of the questionnaire can be found in machine-readable formats on GitHub,¹⁰ and can be applied to any PGIS. To ensure maximum re-usability in research and industry alike, PGUS is released as Open Data under a Creative Commons Attribution Share-Alike 4.0 (CC-BY-SA-4.0).

6 A Case Study on SeaSketch

The PGUS was implemented and disseminated to the SeaSketch user base, in order to probe the usability of the platform along multiple dimensions. For the purpose of this study, the PGUS was complemented by demographic and project-specific questions, aimed at understanding the composition of the SeaSketch user base and their self-assessed level of expertise. For the sake of completeness, we include these additional ten questions in Table 2, which might be used as a template in other PGIS projects.

The survey was disseminated by email to the SeaSketch user community through the SurveyMonkey web platform.¹¹ The invitation was delivered to a set of 3,274 user emails, including inactive and occasional users. These calls obtained 288 anonymous responses, followed by a tail of 23 late responses, for a total of 311 responses. Out of all responses, 59% were complete, while the other respondents did not complete the questionnaire. The total number of complete responses is therefore 181, corresponding to 6% of the whole user base.

Completion Time

The PGUS is designed to maximize user participation and completion time is therefore an important dimension to observe. In the set of complete responses, we calculated the minutes elapsed between the beginning and the end of the user sessions. This resulted in 181 completion times, with the following distribution: minimum (0.8 min), 1st quartile (2.8 min), median (4.2 min), 3rd quartile (7.4 min), and maximum (2,611 min). Note that these times include the ten questions about the user, which are not part of the PGUS.

Hence, based on the median of 4.2 min, we can confirm that the completion time for PGUS is well below 5 min. In this distribution, the tail of low completion times is of particular interest to detect invalid responses. Considering the length and complexity of the questionnaire, we estimated that responses that took less than 90 s were generated without comprehending its content. Six responses were therefore discarded, leaving 175 valid responses in the dataset.

¹⁰<http://github.com/andrea-ballatore/pgis-usability>.

¹¹The survey was disseminated on October 19, 2015, with a reminder on October 26.

Table 2 The user questionnaire that was designed to collect information about the demography and familiarity with SeaSketch of participants. Note that this is not part of PGUS

	Welcome to the SeaSketch User Survey! We are the SeaSketch team, and we are collecting feedback to improve SeaSketch. This survey is strictly anonymous, and it will take less than 5 minutes to complete it. Please note that this survey only applies if you used SeaSketch. If you have any questions about this survey, please contact us at <email>.
Q1	What is your age? <17 or younger; 18–20; 21–29; 30–39; 40–49; 50–59; 60 or older>
Q2	Are you male or female? <Male; Female>
Q3	What is the highest level of education you have completed? < Did not attend school, ..., Completed graduate school>
Q4	How would you rate your web experience prior to working with SeaSketch? <Beginner; Intermediate; Expert>
Q5	How long have you used SeaSketch for? <1–10h; 11–20 h; 21 or more h>
Q6	How would you rate your knowledge of SeaSketch? <Beginner; Intermediate; Expert>
Q7	Have you ever contributed to other public participatory projects? <Yes; No>
Q8	What problems do you encounter using SeaSketch? <Open answer>
Q9	What existing features of SeaSketch need improving? <Open answer>
Q10	What new features would you like SeaSketch to offer? <Open answer>

Demographics

Out of 175 respondents, 97 males (55%) and 78 females (45%). The age of respondents is illustrated in Fig. 2, showing that 26% are in the 30–39 group, followed by 24% in 40–49 group. Another demographic dimension is the level of education, showed in Fig. 3. Surprisingly, 60% of respondents completed graduate school, and 22% graduated from college, indicating a very high level of education of SeaSketch users, considerably higher than other PGIS projects.

Two dimensions aim at capturing the level of experience of users with the tool, and the total number of hours of prior usage. As shown in Fig. 4, 25% of respondents considered themselves as beginners, followed by a 40% of intermediate, and a 35% of experts. This self-assessment is consistent with the number of estimated hours of exposure to the tool. Apart from the level of education, heavily skewed towards the highest part of the range, the sample of 175 appears to cover in a balanced way males and females, as well as a broad range of age groups and levels of expertise.

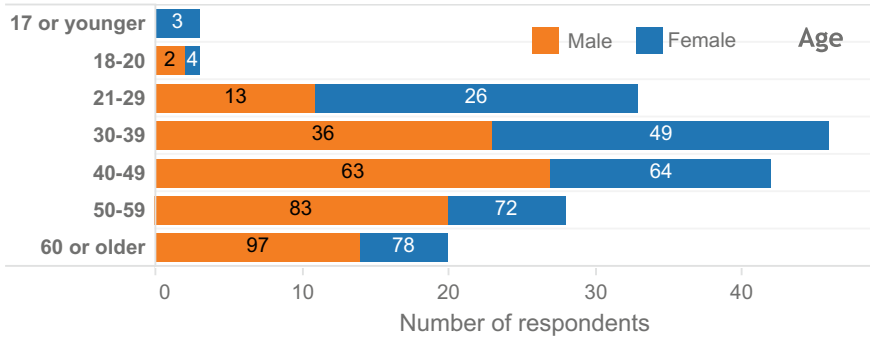


Fig. 2 Age of the respondents, including number of males and females (N = 175)

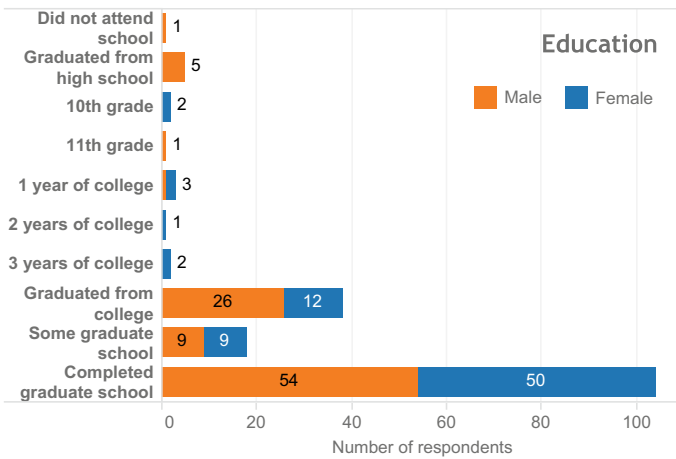


Fig. 3 Level of education of the respondents, including number of males and females (N = 175)

Tool Expertise

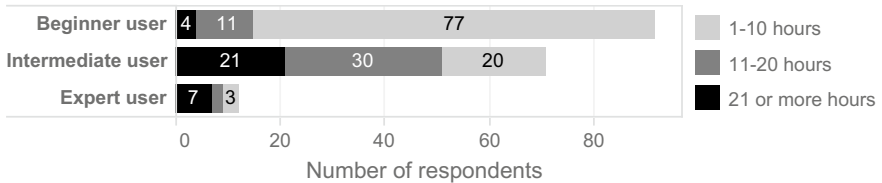


Fig. 4 Self-assessed expertise of the respondents (N = 175)

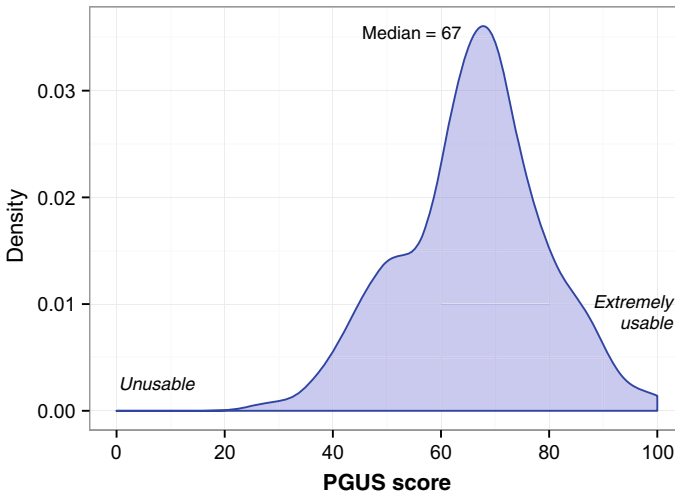


Fig. 5 Density of PGUS scores (N = 175)

PGUS Scores

The ratings in the Likert scales were used to calculate 175 usability scores, one for each respondent. The distribution of these scores goes from a minimum of 28 to a maximum of 100, with the first quartile being 59, and the third quartile being 74. More importantly, the median PGUS score is 67. Figure 5 shows the density of the scores, highlighting the median. Subsequently, we analyzed scores to find correlations with the demographic characteristics of respondents. Using the non-parametric Kruskal-Wallis rank sum test (Sprenst and Smeeton 2007), no significant difference was detected between scores by males and females.

The age of respondents has a small effect on the variance of the PGUS scores. The Kruskal-Wallis test detects significant differences among the age groups (chi-squared 17.8 on 6 degrees of freedom, with $p < 0.01$). Treating age groups and PGUS scores as ordinal variables, a weak inverse correlation can be observed, indicating that young users tend to find the system more usable than older users (Spearman's $\rho = -0.25$, with $p < 0.01$). No other significant correlations were identified between PGUS scores and the self-assessed level of expertise and hours of exposure to the tool.

PGUS Likert Ratings

After discussing the PGUS scores, which captures the overall perception of usability, it is useful to consider the detailed Likert ratings to the 25 questions that are used in the score calculation. Overall, the 175 respondents expressed 4,375 ratings on the Likert scale. The counts of ratings for each question of the PGUS are displayed in Fig. 6. The figure shows the sum of the PGUS scores only for each specific question, with a distance from the mean expressed as a percentage. For example, question D4 (“The system increases my participation in the project”) obtains score 462, slightly

is below average (-4.2%). Based on these considerations, it is possible to state that the variation across questions and groups is contained between -12.2% and 15.1% , without showing any extreme variation. The variation is even lower when observing the five themes, ranging from -4.2% to 3.9% . As suggested by Brooke (2013), uniformity of ratings is a further indicator of the soundness of the questionnaire.

Limitations

This exploration of SeaSketch with PGUS was useful to the development team, revealing different facets of the usability of the tool. However, this initial version of the questionnaire has important limitations that should be borne in mind. This questionnaire should be analyzed in terms of validity (the measurement is accurate) and reliability (the measurement is consistent). Usability judgments are subjective by definition, as opposed to performance metrics that can be assessed more objectively (e.g., execution time of a task). As a result, the validity of PGUS is a complex issue, and we assessed it with a qualitative *content validity* assessment, based on discussions with a sample of participants from the SeaSketch community to ensure the relevance and the clarity of questions. As we did not have access to a PGIS that is well-known for poor usability, we could not assess validity with comparative methods. To assess PGUS reliability, we have observed good internal consistency. Repeating the assessment over time on different platforms is the ideal way further ensuring reliability. For both validity and reliability, we hope that the research community will use PGUS in diverse contexts and share the results, generating insights and possible improvements that are hard to identify within a single PGIS project.

The meaning of the PGUS score is at the moment only interpretable in relative terms, for example stating how the judgment of two users differ and whether a facet is perceived as more or less usable than another one. Even in well-established surveys like SUS, it is not trivial to translate the 0–100 scores into meaningful adjectives, such as “poor” and “excellent,” establishing general guidelines, e.g., scores lower than 50 indicate unacceptable usability (Bangor et al. 2008). Applying PGUS to other systems appears necessary to gather more knowledge and ascertain typical scores for practitioners.

While PGUS aims at being broadly applicable to any PGUS, some questions are likely to be more relevant than others to specific problems. For example, “I can create new content easily” might not apply to visualization-only tools. In this sense, there is a tension between the advantages of a standardized questionnaire and ad-hoc solutions that will better fit the specificities of a project. Other limitations concern the thematic and semantic scope of PGUS. Some questions are about the content of the platform, rather than just about the user interface (“the maps are easy to understand”). For example, users might create poor quality spatial content on a very usable platform, and vice-versa. More conceptual separation between these aspects would be beneficial. Similarly, the questionnaire includes aspects of usability (e.g., B3) and user experience (e.g., E5). In future versions, it might be beneficial to group those aspects more clearly.

7 Conclusion

PGIS is an important application area for GIS and spatial technologies, with specific challenges and user needs. Despite the remarkable advances in usability in general, it can be argued that PGIS usability research has provided limited insights to developers. In this paper, we have argued that the lean approach to usability is particularly suitable for PGIS, and we have started to outline a set of techniques to embed it into an existing platform. As a case study, we focused on SeaSketch, a web platform for participatory marine spatial planning, currently adopted in numerous projects around the world.

To support usability in PGIS, we proposed to combine heuristic usability evaluations, user feedback buttons, intercept surveys, email surveys, and implicit feedback. These techniques provide multi-modal, flexible probes to collect opinions, with the purpose of identifying usability issues. We expect the deployment of this ensemble of lean techniques in SeaSketch to rapidly uncover usability issues, leading towards the development of comprehensive PGIS usability guidelines, updated to the current technological landscape.

As a first step towards this framework, we designed the Participatory GIS Usability Scale (PGUS), providing a simple and inexpensive mechanism to measure the perceived usability of a PGIS. The PGUS was designed in collaboration with the SeaSketch community, starting from Brooke's SUS, a widely used usability questionnaire (Brooke 2013). This questionnaire contains 25 questions, and quantifies the perceived usability of a system on a scale between 0 and 100. The PGUS was disseminated to the SeaSketch user base, obtaining 175 complete responses. The analysis of the results highlights the perceived strengths and weaknesses of the tool, providing useful indications to the developers.

This article covers the first version of the PGUS, which provides a useful starting point for PGIS practitioners. However, more research and evaluations are needed to further refine the questionnaire, assessing its validity and reliability across different PGIS communities. For example, formal validations of the PGUS can be obtained by measuring correlations with more established questionnaires (Hornbæk 2006), as well as comparing the scores of systems that have known excellent and poor usability.

Of the many challenges in PGIS usability, better information sharing mechanisms and the personalization of the interface for users from different backgrounds are of particular interest (Ballatore and Bertolotto 2015). Moreover, PGIS usability cannot ignore recent trends in web design, such as responsive design and the mobile-first strategy, as they are likely to have a broad impact in the near future of PGIS. Understanding what works and what does not on tablets and smartphones is paramount in the current technical landscape (Nielsen and Budiu 2013). Notably, touch interfaces constitute growing and under-explored media to lower the barriers to participatory spatial systems (Haklay 2010).

Because of its civil and political importance, the future PGIS deserves not only usability evaluation for the sake of efficiency and effectiveness, but better user experience from social and affective viewpoints, reducing the need for skilled facilitators

(the “GIS chauffeurs”) that are still playing a major role in many projects. The limited empirical evidence and insights that we possess invites further research for GIScientists, software developers, social scientists, and planners to study usability in PGIS. Although usability itself cannot overcome deep social and institutional barriers to participation, more usable tools are essential to maximize the inclusiveness, accessibility, and ultimately the likelihood of success of participatory projects.

Acknowledgements We thank Noah Gluschankoff (University of California, Santa Barbara) for his work on the questionnaire, and the SeaSketch users for their detailed feedback on the early drafts. Initial funding and support for SeaSketch have been provided by Esri, the New Zealand Department of Conservation, and The Tindall Foundation.

References

- Aditya T (2010) Usability issues in applying participatory mapping for neighborhood infrastructure planning. *Trans GIS* 14(s1):119–147
- Ballatore A, Bertolotto M (2011) Semantically enriching VGI in support of implicit feedback analysis. In: Tanaka K, Fröhlich P, Kim K-S (eds.) *Proceedings of the web and wireless geographical information systems international symposium, LNCS, vol 6574*. Springer, Berlin, pp 78–93
- Ballatore A, Bertolotto M (2015) Personalizing maps. *Commun ACM* 58(12):68–74
- Balram S, Dragičević S (2005) Attitudes toward urban green spaces: Integrating questionnaire survey and collaborative GIS techniques to improve attitude measurements. *Landsc Urban Plan* 71(2–4):147–162
- Bangor A, Kortum PT, Miller JT (2008) An empirical evaluation of the system usability scale. *Int J Hum Comput Interact* 24(6):574–594
- Brooke J (1996) SUS: a ‘quick and dirty’ usability scale. In: Jordan PW, Thomas B, McClelland IL, Weerdmeester B (eds) *Usability evaluation in industry*. CRC Press, Boca Raton, FL, pp 189–194
- Brooke J (2013) SUS: a retrospective. *J Usability Stud* 8(2):29–40
- Brown G (2012) Public participation GIS (PPGIS) for regional and environmental planning: reflections on a decade of empirical research. *J Urban Reg Inf Syst Assoc* 25(2):7–18
- Brown G, Kytä M (2018) Key issues and priorities in participatory mapping: toward integration or increased specialization? *Appl Geogr* 95:1–8
- Brown G, Sanders S, Reed P (2018) Using public participatory mapping to inform general land use planning and zoning. *Landsc Urban Plan* 177(April):64–74
- Brown G, Strickland-Munro J, Kobryn H, Moore SA (2017) Mixed methods participatory GIS: an evaluation of the validity of qualitative and quantitative mapping methods. *Appl Geogr* 79:153–166
- Dunn CE (2007) Participatory GIS—A people’s GIS? *Prog Hum Geogr* 31(5):616–637
- Ehler C, Douvère F (2009) *Marine spatial planning: a step-by-step approach toward ecosystem-based management*. Technical report, UNESCO, Paris
- Evans AJ, Kingston R, Carver S (2004) Democratic input into the nuclear waste disposal problem: the influence of geographical data on decision making examined through a web-based GIS. *J Geogr Syst* 6(2):117–132
- Forrester J, Cinderby S (2013) A guide to using community mapping and participatory-GIS. <https://web.archive.org/web/20180417144258/>, http://tweedforum.org/research/Borderlands_Community_Mapping_Guide_.pdf
- Frøkjær E, Hertzum M, Hornbæk K (2000) Measuring usability: are effectiveness, efficiency, and satisfaction really correlated?. In: *Proceedings of the SIGCHI conference on human factors in computing systems*, ACM, The Hague, Amsterdam, pp 345–352

- Gardner BS (2011) Responsive web design: enriching the user experience. *Sigma J: Digit Ecosyst* 11(1):13–19
- Garrett JJ (2010) *The elements of user experience: user-centered design for the web and beyond*, 2nd edn. Pearson Education, New York
- Gleason M, McCreary S, Miller-Henson M, Ugoretz J, Fox E, Merrifield M, McClintock W, Serpa P, Hoffman K (2010) Science-based and stakeholder-driven marine protected area network planning: a successful case study from north central California. *Ocean Coast Manag* 53(2):52–68
- Goldberg G, Diorio M, McClintock W (2016) *Applied marine management with volunteered geographic information*. CRC Press, Boca Raton, FL, pp 149–174
- Goodchild MF (2010) Towards geodesign: repurposing cartography and GIS? *Cartogr Perspect* 66:7–22
- Haklay M (ed) (2010) *Interacting with geospatial technologies*. Wiley, Hoboken, NJ
- Haklay M, Singleton A, Parker C (2008) Web mapping 2.0: The neogeography of the GeoWeb. *Geogr Compass* 2(6):2011–2039
- Haklay M, Tobón C (2003) Usability evaluation and PPGIS: towards a user-centred design approach. *Int J Geogr Inf Sci* 17(6):577–592
- Hassenzahl M, Tractinsky N (2006) User experience - a research agenda. *Behav Inf Technol* 25(2):91–97
- Hornbæk K (2006) Current practice in measuring usability: challenges to usability studies and research. *Int J Hum Comput Stud* 64(2):79–102
- Krug S (2014) *Don't make me think, revisited: a common sense approach to web usability*. New Riders, San Francisco, CA
- Kwaku Kyem PA (2004) Of intractable conflicts and participatory GIS applications: the search for consensus amidst competing claims and institutional demands. *Ann Assoc Am Geogr* 94(1):37–57
- Lewis JR (1995) IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *Int J Hum Comput Interact* 7(1):57–78
- Maquil V, Leopold U, De Sousa LM, Schwartz L, Tobias E (2018) Towards a framework for geospatial tangible user interfaces in collaborative urban planning. *J Geogr Syst* 20:185–206
- Mare Nostrum (2016) *Final report: legal-institutional instruments for integrated coastal zone management (ICZM) in the Mediterranean*. Technical report, Mare Nostrum Partnership, European union
- Merrifield MS, McClintock W, Burt C, Fox E, Serpa P, Steinback C, Gleason M (2013) *MarineMap: a web-based platform for collaborative marine protected area planning*. *Ocean Coast Manag* 74:67–76
- Nielsen J (1999) *Designing web usability: the practice of simplicity*. New Riders, Berkeley, CA
- Nielsen J (2004) Risks of quantitative studies. Nielsen Norman Group. <http://www.nngroup.com/articles/risks-of-quantitative-studies>
- Nielsen J (2009) Discount usability: 20 years. Nielsen Norman Group. <http://www.nngroup.com/articles/discount-usability-20-years>
- Nielsen J (2012a) Usability 101: introduction to usability. Nielsen Norman Group. <http://www.nngroup.com/articles/usability-101-introduction-to-usability>
- Nielsen J (2012b) User satisfaction vs. performance metrics. Nielsen Norman Group. <https://www.nngroup.com/articles/satisfaction-vs-performance-metrics>
- Nielsen J, Budiu R (2013) *Mobile usability*. New Riders, Berkeley, CA
- Nivala A-M, Brewster S, Sarjakoski TL (2008) Usability evaluation of web mapping sites. *Cartogr J* 45(2):129–138
- Norman DA (2013) *The design of everyday things (revised and expanded edition)*. Basic Books, New York
- Poplin A (2012) Playful public participation in urban planning: a case study for online serious games. *Comput Environ Urban Syst* 36(3):195–206
- Raubal M (2009) Cognitive engineering for geographic information science. *Geogr Compass* 3(3):1087–1104

- Rinner C, Keßler C, Andrulis S (2008) The use of Web 2.0 concepts to support deliberation in spatial decision-making. *Comput Environ Urban Syst* 32(5):386–395
- Rohrer C (2008) When to use which user experience research methods. Nielsen Norman Group. <http://www.nngroup.com/articles/which-ux-research-methods>
- Roth RE (2015) Interactive maps: what we know and what we need to know. *J Spat Inf Sci* 6:59–115
- Sauro J (2011) A practical guide to the system usability scale: background, benchmarks & best practices. CreateSpace, Seattle, WA
- Schlossberg M, Shuford E (2005) Delineating “public” and “participation” in PPGIS. *URISA J* 16(2):15–26
- Sieber R (2006) Public participation geographic information systems: a literature review and framework. *Ann Assoc Am Geogr* 96(3):491–507
- Skarlatidou A, Cheng T, Haklay M (2013) Guidelines for trust interface design for public engagement web GIS. *Int J Geogr Inf Sci* 27(8):1668–1687
- Slocum TA, Blok C, Jiang B, Koussoulakou A, Montello DR, Fuhrmann S, Hedley NR (2001) Cognitive and usability issues in geovisualization. *Cartogr Geogr Inf Sci* 28(1):61–75
- Smith DA (2016) Online interactive thematic mapping: applications and techniques for socio-economic research. *Comput Environ Urban Syst* 57:106–117
- Sprent P, Smeeton NC (2007) Applied nonparametric statistical methods. CRC Press, Boca Raton, FL
- Stevens M, Vitos M, Altenbuchner J, Conquest G, Lewis J, Haklay M (2014) Taking participatory citizen science to extremes. *IEEE Pervasive Comput* 13(2):20–29

Future Occupational Profiles in Earth Observation and Geoinformation—Scenarios Resulting from Changing Workflows



Barbara Hofer, Stefan Lang and Nicole Ferber

Abstract Technological advances require continuous efforts to keep existing curricula up-to-date and graduates employable in the Earth observation (EO) and geoinformation (GI) sectors. The increasing availability of space/geospatial data and the maturity of technology induce disruptive changes to workflows in the EO/GI sector that suggest the development of training programmes and academic courses for re-skilling of workforce and training new user groups. The target in the EO domain in this respect is to facilitate the ‘user uptake’ of the space infrastructure. User uptake requires knowledge of the workforce demand on the market as well as a skills strategy that takes potential emerging and disruptive changes in the sector into account. In the present contribution we build upon a study of demand for current workforce on the EO/GI market and occupational profiles that require priority when developing training programmes and curricula. Reflections on the findings of that study highlight the need to illustrate expected changes of workflows, i.e. the sequence of tasks executed by employees with a certain occupational profile, for an improved basis of discussion. Therefore, we present a methodology to first, acquire current occupational profiles and second, to illustrate sector developments by mapping the developments on tasks of the workflow. This methodology is demonstrated for the profile of remote sensing specialists. The illustration of changing tasks suggests scenarios for future workforce and questions and directions for the development of a sector skills strategy.

Keywords Earth observation · Geoinformation · User uptake · Sector skills strategy · Technological trends

B. Hofer (✉) · S. Lang · N. Ferber
Interfaculty Department of Geoinformatics – Z_GIS, Paris-Lodron-University of Salzburg,
Schillerstr. 30, 5020 Salzburg, Austria
e-mail: barbara.hofer@sbg.ac.at

S. Lang
e-mail: stefan.lang@sbg.ac.at

N. Ferber
e-mail: nicole.ferber@sbg.ac.at

© Springer Nature Switzerland AG 2020
P. Kyriakidis et al. (eds.), *Geospatial Technologies for Local and Regional
Development*, Lecture Notes in Geoinformation and Cartography,
https://doi.org/10.1007/978-3-030-14745-7_19

1 Introduction

The geoinformation (GI) and in particular the Earth observation (EO) sectors are currently experiencing disruptive changes because technological developments are mature enough to enable automation and policies strongly support digitization paradigm. The developments are widely discussed in literature (Chen et al. 2016; Li et al. 2016; O’Sullivan et al. 2018; Sudmanns et al. 2018). What requires further discussion is the effects of these developments on current workflows, whereby a workflow is the sequence of tasks executed by employees with a certain occupational profile (with occupational profile or job profile we refer to the knowledge, skills and competencies required for fulfilling tasks related to a specified job).

The need to identify future occupational profiles that support user uptake and the development of training programmes and curricula for such profiles has been identified in preparation of the EO4GEO project. ‘EO4GEO—Towards an innovative strategy for skills development and capacity building in the space geo-information sector supporting Copernicus User Uptake’, is an Erasmus + Sector Skills Alliance project with 26 partners from the public and private sectors and academia from 13 European Union member states (<http://www.eo4geo.eu>). The project started in January 2018 and has a run time of four years. Current work tasks in the project concern the analysis of supply and demand of workforce on the EO/GI market, the development of an integrated EO/GI Body of Knowledge and the conception of an innovative strategy for skills development and capacity building.

As expressed in the project name, EO4GEO aims at developing a joint strategy for the EO and the GI sectors. In the past the EO and GI sectors were mostly understood to operate in parallel. However, the development of information products for the variety of application areas of space and geospatial data requires an integration of data created in both of the sectors. The authors claim that at the core of both sectors is an understanding for the nature of spatial data and technology and how they can be exploited in products and services. Visions of the future foresee a full integration of EO and GI data in spatial data infrastructures (Lang et al. 2018). The purpose of updating geospatial data may be regarded as one of the constant future requirements, irrespective of the specific application context, the pace (regular update intervals or quasi-continuous), and processing platform (cloud-based or even on-board).

The 2017 call for proposals for the Erasmus + programme of the European Commission (EC) requested a skill strategy for the *space geospatial sector*. This request is connected to activities of the Copernicus programme—the joint Earth observation (EO) programme by the EC and the European Space Agency (ESA). Therefore, it is timely to jointly consider the EO and GI sectors when developing a skill strategy with the objective of user uptake.

In preparation of the sector skill strategy, EO4GEO analyses the supply and demand for skills, knowledge and competencies of the (current) workforce. The gap between supply (i.e. existing training programmes and courses on various educational levels) and demand is supposed to steer the design of training programmes and curricula as well as training content for future workforce. The insights gained

in the study for assessing demand for workforce mostly depict which skills and occupational profiles currently are required on the EO/GI market. Given the implicit challenge of curriculum development to prepare students/trainees for the future, the present contribution suggests an extension of the previous findings with an analysis of changing workflows. Specifically, we visualize expected changes due to data-driven and technological advances on the existing occupational profile of a remote sensing specialist.

To realize the objective of depicting future occupational profiles, this contribution builds on principles of technology forecasts to provide an improved basis of discussion for the future development of an EO/GI sector skill strategy (Committee on Forecasting Future Disruptive Technologies 2010). An intuitive method based on expert knowledge is applied for forecasting changes in the sector that affect workforce demands. The Committee on Forecasting Future Disruptive Technologies (2010) differentiates three actions of technology forecasts: (1) definition of forecast objectives, (2) data gathering and analysis, (3) result interpretation.

The objective of the targeted forecast is the depiction of the impact of technological advances on duties and tasks of occupational profiles in the sector. For framing the discussion in this contribution, the profile of a remote sensing (RS) specialist is selected. Data gathering and analysis builds on the one hand on the design of a curriculum method (DACUM) (Norton 1997) and on the other hand on the identification of developments and trends that will impact the work of remote sensing specialists. Result interpretation focuses on the identification of directions of developments and questions for strategy development. The time frame of the targeted forecast is medium term, i.e. envisaged changes within five to ten years. As this contribution is based on a forecast of changes, we are aware that the discussion relates to one possible vision of the future.

In Sect. 2, the findings of the study on skills demand of workforce are summarized and critically reflected. Section 3 introduces the design a curriculum method (DACUM) that leads to the specific outline of duties and tasks of remote sensing specialists as a (current) occupational profile. In preparation of the suggested workflow analysis with particular focus on this profile, Sect. 4 presents disruptive changes on the EO sector. Section 5 discusses changes of the remote sensing workflow that are introduced by the technological developments presented in Sect. 4. The questions that arise from the analysis of a changing workflow for the development of a sector skill strategy are discussed in Sect. 6.

2 EO4GEO Analysis of Demand for EO/GI Skills

Information about required knowledge, skills and competencies to satisfy occupational profiles are a key for training the future workforce (Cedefop 2013). The European Centre for the Development of Vocational Training (Cedefop) provides the following definitions of knowledge, skills and competence:

- Knowledge: “Outcome of assimilation through learning. Knowledge is the body of facts, principles, theories and practices related to a field of study or work” (Cedefop 2014, p. 147).
- Skills: “Ability to apply knowledge and use know-how to complete tasks and solve problems” (Cedefop 2014, p. 227).
- Competence: “Ability to use knowledge, skills and personal, social and/or methodological abilities, in work or study situations and in professional and personal development” (Cedefop 2014, p. 47).

The analysis of the current demand for EO/GI occupational profiles in the EO4GEO project consisted of several components: an online survey, semi-structured interviews and a workshop for discussing preliminary findings with market representatives (Aguilar Moreno et al. 2018; Hofer et al. 2018).

The starting point for the analysis was the online survey, which can be considered a follow-up survey to the GI-focused survey in the GI-N2K project (Wallentin et al. 2015). The main aim of the survey was to acquire descriptions of occupational profiles that are currently in high demand on the EO/GI market and allow deriving priority occupational profiles for the subsequent development of training programmes. For the specification of these occupational profiles, sets of skills were presented to the respondents. These sets of skills resemble knowledge areas of the GIS&T Body of Knowledge (BoK) and cover: space/geospatial data, data capture and management, analytical methods, programming and development, computing resources and platforms, visualization and cartography, organizational and institutional issues, EO/GI and society. The respondents were presented with a series of examples of elements for each skill set (e.g., *extraction, transformation and loading of EO/GI data* and *orthorectification and mosaicking of EO data* were two of 14 examples of the space/geospatial skill set); they then provided an overall rating of each skill set between 1 (not relevant) and 6 (extremely relevant).

In the time between April and July 2018, 175 responses to the online survey were collected mostly from across Europe.¹ The set of skills that stands out across the specified profiles is *space/geospatial data skills* (Fig. 1). The four sets of skills that follow the data skills (*visualization and cartography, analytical methods, programming and development, data capture and management*) seem nearly equally important. The last three sets of skills (*EO/GI and society, computational resources and platforms and organizational and institutional aspects*) are the ones rated least important. This aggregated picture of the skill ratings was differentiated by organization type and provided an indication that different types of organizations require employees with different skills: *programming and development* was the most important skill set in occupational profiles described as important in large companies; in contrast, for public organizations there was little divergence from the overall average presented in Fig. 1. In the education and vocational training context, the skill set

¹The geographical distribution of collected responses is available here (January 2019): https://public.tableau.com/views/EO4GEO_Demand_Survey_Part1_Overview_FA/MapDashboard?:embed=y&:display_count=yes&publish=yes.

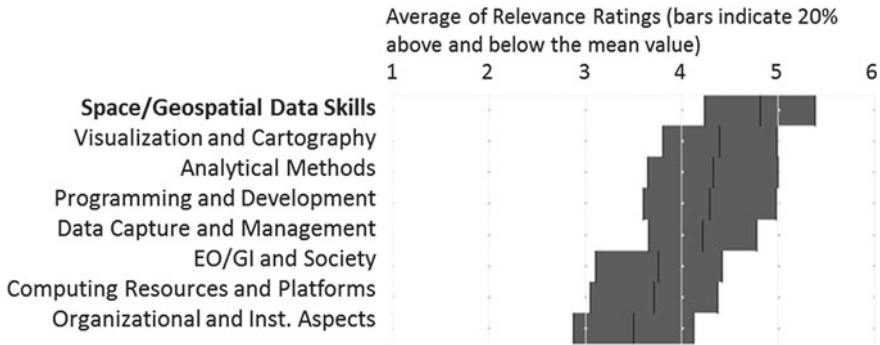


Fig. 1 The overall relevance of the EO/GI sets of skills, ordered by mean relevance rating (mean indicated by vertical black lines)

analysis methods was rated more important, but still ranks behind data skills (details in Hofer et al. 2018).

The 175 occupational profiles collected in the survey are heterogeneous regarding the ratings of sets of skills for specific occupational profiles. Examples of skill set ratings of the profile of a *GIS developer* on Master level are shown in Fig. 2.

There is little agreement on the importance of skill sets in profiles with the same or similar name, which means that general labels like *GIS&T specialist* or *RS specialist* subsume occupational profiles with different emphasis.

Based on the overall ratings of skills sets and on the overlap between described occupational profiles, three occupational profiles were identified as an outcome of the survey: *EO/GI developer*, *EO/GI data analyst* and *EO/GI project manager* (Hofer et al. 2018).

In contrast to the specific skills requested, there seems to be agreement on the training level for occupational profiles in the EO/GI market: virtually all requested profiles are on a master level (52%) or PhD level (34%). Only three profiles have been described for people who are high school graduates or went through vocational training. These figures correspond to the characterization of current workforce in the survey that indicates that 85% have a master or Ph.D. degree themselves.

The semi-structured interviews with 30 market representatives of the EO/GI sector confirmed that the specific relevance of specific skills, knowledge and competencies for an occupational profile depends on factors like the type and size of organisations, the business processes and the specific tasks of an individual. A general point mentioned in the interviews was that with the rise of new technologies the organization of work changes as well as the flexibility and fragmentation of activities and workflows. Furthermore, a shift from classical remote sensing skills towards programming, cloud computing, harvesting, and understanding computing systems was mentioned.

The results of the demand analysis need to be critically reflected:

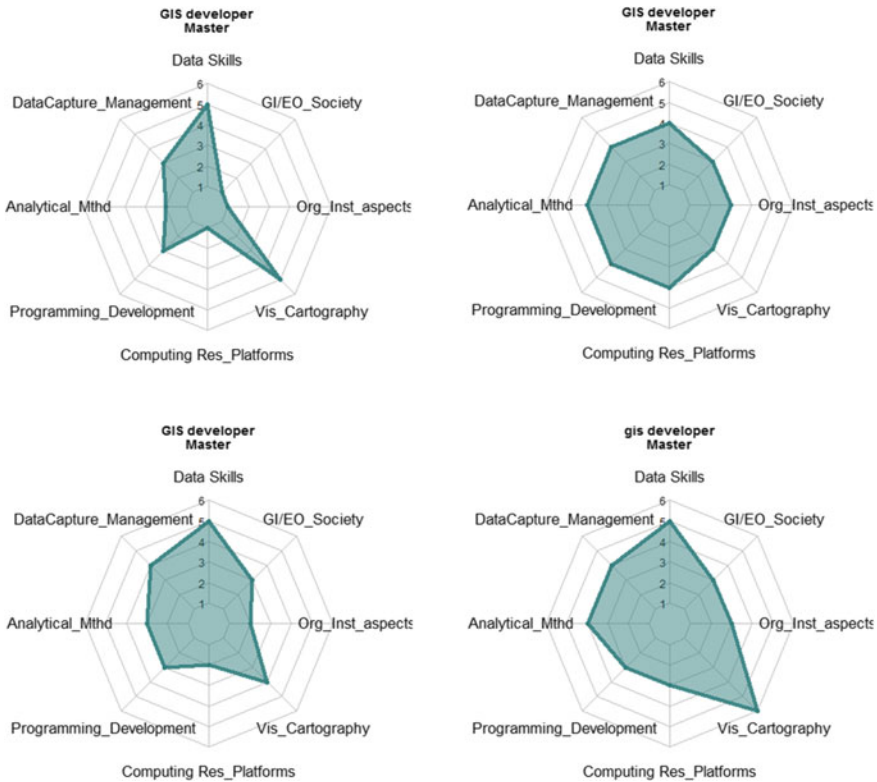


Fig. 2 Four specific skill set ratings of distinct respondents of the survey; despite the difference in the ratings all of them were attribute to the profile of a GIS developer

- The results of the demand analysis represent current demand on the market and include little anticipation of the future demand in response to changing workflows.
- The specification of occupational profiles based on skill sets used in the survey might exclude competencies that are required in a daily work context.
- Respondents were mostly from the EO/GI sector: the responses tend to reflect the opinions of representatives who are active in the field. Thus, a more ‘traditional’ view on the matter is created than potential newcomers (so-called new-space entrants or non-space industries) may have. While also spin-off companies and start-ups are included in the group of recipients of the survey, the number of unconventional views from outside the remote sensing community is limited.

Given the findings of the demand analysis, we observed that it is not enough to focus on what is currently requested on the market as there is a tendency to request profiles that resemble the existing ones. Future user uptake might be facilitated by vocational training programmes and certificates that are seemingly not currently requested on the market. These observations are translated into a workflow oriented

analysis of future occupational profiles that has the potential to better differentiate requested profiles.

3 Current Workflow of Remote Sensing Specialists

For the provision of a more detailed basis of discussion of future occupational profiles, we specifically focus on the forecasting of changes of the profile of a remote sensing (RS) specialist. This section reports on the characterisation of a current RS specialist profile and the workflow associated with this profile.

We followed the ‘developing a curriculum’ (DACUM) method for the identification of duties and tasks of a RS specialist (Collum 1999; Norton 1997; Tippelt and Edelmann 2007). DACUM has been designed in the 1960s in Canada for achieving a job-oriented education (Kunz 2015). The aim of DACUM is to analyse a job and to derive duties and tasks as well as required technical skills and personal traits in recurring discussion rounds. The discussions involve experts in the field who reach consensus on the job profile during the discussion process. The selection of experts for the DACUM discussions is thereby essential for the quality of the outcome. There are a series of examples of DACUM profiles of the GI and EO sectors (<http://www.geotechcenter.org/gtcm-dacum.html>) as well as a meta-analysis of developed profiles in the GI sector (Johnson 2010).

Although the method name refers to curriculum development, the foundation of this method is the analysis of tasks and duties associated with a specific job profile. Taking tasks and duties of a workflow as a starting point assures that the subsequently derived content of curricula fulfils the job-related requirements.

For the purpose of this contribution, we applied a slightly downgraded variation of the DACUM method: we brought together four remote sensing specialists of the Department of Geoinformatics for a half-day DACUM workshop. The number of four experts is a little lower than the suggested 6–12 experts (Tippelt and Edelmann 2007). The four invited experts cover a variety of different focus areas in remote sensing and work on service provision rather than research tasks despite the fact that they are employed at a university department. The focus in our DACUM workshop was on duties and tasks; complementary information on personal characteristics, software skills etc. was not discussed in detail. As each of the experts is responsible for the whole workflow in their area of expertise, the resulting profile is universal in terms of the tasks that the RS specialist has to perform. There are some specific properties of the occupational profile under discussion: communication is seen as a separate duty, which gives importance to competencies in this part of the work. Also, data do not only include EO satellite data, but also LiDAR data and other geospatial datasets that are created or used as part of the workflow.

Profiles created by different groups of experts might differ from the one created in this specific workshop; however, the applied DACUM method is supposed to provide sufficient insight in the work of an RS specialist to support the illustration of changes of tasks and duties in the remaining sections of this paper.

The chart in Fig. 3 shows the duties and tasks of a remote sensing specialist together with a categorization of the tasks according to the prevailing skills or competencies required for the tasks. The rating of skills has again been performed by experts and differentiates three categories: tasks requiring transversal skills, technical skills or competencies in the EO domain, whereby competencies are the combined use of knowledge and skills for tasks together with personal and social abilities and also include experience (Cedefop 2014). For tasks requiring technical skills the experts assumed that standardized procedures or guidelines on how to perform tasks exist or algorithms are in place to support the specialists. For example, the task of geometric correction, which is here classified as task requiring technical skill, might require competencies if there are no routines available for the correction. The categorization of tasks gives an indication of which tasks can be assigned to less experienced workers and which tasks certainly require a high skill level and experience in the RS domain.

4 Technology Trends in the Earth Observation Sector

The collection of imagery from satellites or other sensor systems has the aim to produce space-time information and knowledge for decision making (Lang et al. 2018). We argue that the ways in which imagery is collected and exploited is undergoing disruptive changes at the moment. The disruptive nature of changes is due to the interplay and maturity of developments like cloud infrastructures and big EO data (Chen et al. 2016; Lang et al. 2018; Li et al. 2016; O’Sullivan et al. 2018). This section summarizes the main changes of the EO sector that impact the work of EO specialists.

The value chain of the EO sector, covering upstream, midstream and downstream services, has constantly been evolving (Fig. 4). Both the upstream sector, i.e. the (space) infrastructure with its massive increase in EO sensors/satellites, and the midstream/downstream sector, i.e. the ground segment plus value-added services, experience significant changes. Space infrastructure is gradually growing, comprising huge, multi-national missions like the Sentinel satellites family, and commercial satellites with increasing spatial resolution, acquisition frequencies, continuous imaging, as well as micro satellites for near-individual, or solitary use.

The ground segment comprising the data storage and access facilities, faces disruptive changes in the way data storing and access is organised, in particular for Sentinel data which is supposed to be full, free and open, and thus increasingly offered as a platform-as-a-service (PaaS). Big EO data or big Earth data (Guo 2017) entails data provision in streams instead of single scenes and an organization of data in data cubes (Sudmanns et al. 2018). The new paradigm of big EO data (“*bring the users to the data*”), entails the provision of analysis-ready data (ARD) in central data infrastructures, and an increase in cloud-based processing and information extraction (Information as a Service, INFOaaS). Concerning image analysis techniques, standardized approaches are sought including machine learning techniques

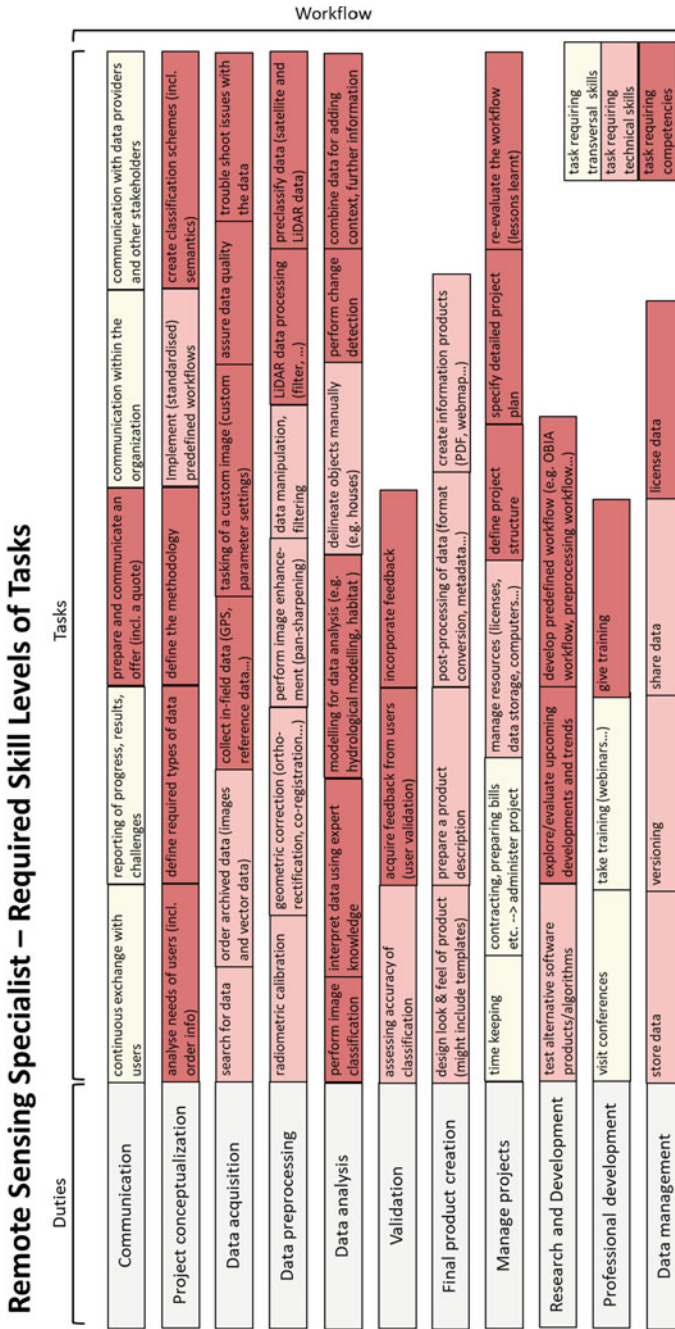


Fig. 3 Occupational profile of a remote sensing specialist generated following the DACUM method

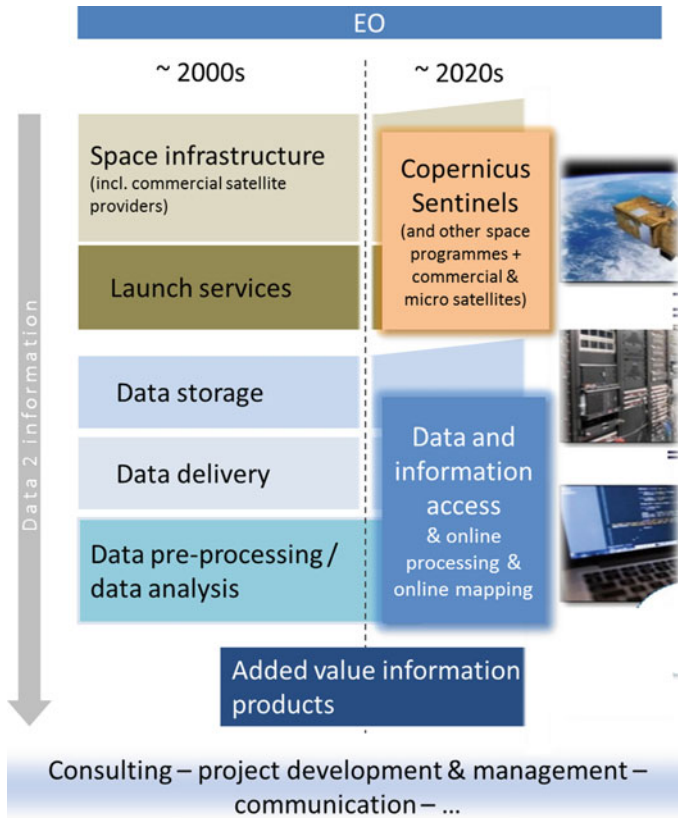


Fig. 4 Evolution of space infrastructure, middleware and product development in the EO sector, based on an illustration of the European Earth Observation Economy (European Commission 2016), modified

like convolutional neural networks (CNN) or automated knowledge-based classification strategies (Lang et al. 2018; Sudmanns et al. 2018). The various implementations of the Copernicus DIAS (Data and Information Access Service), are examples that integrate midstream and downstream services to some degree.

The value-adding industry ('downstream sector') is supposed to benefit enormously from the given data and service infrastructure. These technical achievements need to be ultimately translated into users' speak, i.e. the language and established workflows of users and customers, including the non-EO industry, to understand and recognize technical achievements as related to the challenges they typically face.

The current changes in the EO sector are considered disruptive in the convergence of technological advances (Lang et al. 2018) that occur in each of its components, from data acquisition, access and storage, to image pre-processing and analysis. In addition, a strong integration of EO and GI data in spatial data infrastructures is expected (Lang et al. 2018). This leads to a rapidly growing penetration rate

of automation (automated processes) at all critical elements of the remote sensing workflow, affecting the level and degree of human-machine interaction. The increased automation of the image analysis workflows has consequences for the required skills and the profiles of specialists and users.

In the following chart (Fig. 5) we indicate which tasks of the RS specialist profile are affected by the technological trends defined above:

- the increased use of platforms for data storage (PaaS),
- the availability of analysis ready data (ARD),
- the increased use of platforms for image pre-processing and analysis (INFOaaS),
- the automation of image analysis and the production of scene classification maps (SCM).

The time frame of the forecast is 5–10 years, which means that we expect ARD, PaaS and INFOaaS to have passed their current state of development and to have become part of everyday work of RS specialists.

In this analysis, there is no separate consideration of general developments in the cloud infrastructure that affect also learning and training provision like massive open online courses (MOOCs) or reorganization of the work environment for tasks like collecting user feedback. The analysis specifically focuses on the developments in the EO domain.

For the purpose of this analysis it is assumed that no concerns regarding data security in cloud infrastructures affect the work of the RS specialist. The use of cloud infrastructure can certainly be an issue in some contexts that require security of data. In these cases, the tasks related to the management of the infrastructure cannot be externalised to the provider of the cloud, but remain as task of the RS specialist.

The changes of data handling and EO technology may impact tasks such as discussed in the following exemplary scenarios:

- Tasks affected by PaaS change the execution of the tasks to be based on the interaction with the platform. E.g. there is no need to download Sentinel data, but they can be accessed in an online platform; a further extreme is on-board processing in the space infrastructure itself.
- ARD make tasks obsolete as the data ideally do not require further pre-processing. E.g. radiometric correction does not need to be performed by the RS specialist using the data as this operation has been performed during the preparation of ARD.
- The use of INFOaaS still requires a specialist to operate the tasks, but they are performed in an online environment and with workflows implemented on the platforms. E.g. the creation of classification schemes or applying a standard scheme is an operation supported by the platform that provides the data and no separate software package and desktop tool is required for this operation.
- The automation of image analysis and related operations reduces the resources required for these tasks by the specialists—under the assumption that the tasks are perfectly automated. E.g. image classification relies on standardized approaches that do not require the RS specialist to develop image classification procedures from scratch for every new scene.

Remote Sensing Specialist – Change of Workflow

Duties		Tasks	
Communication	continuous exchange with users reporting of progress, results, challenges	prepare and communicate an offer (incl. a quote)	communication with data providers and other stakeholders
Project conceptualization	analyse needs of users (incl. order info)	define the methodology	create classification schemes (incl. semantics)
Data acquisition	search for data order archived data (images and vector data)	define required types of data (GPS, tasking of a custom image (custom parameter settings))	assure data quality trouble shoot issues with the data
Data preprocessing	radiometric calibration geometric correction (ortho-rectification, co-registration...)	perform image enhancement (pan-sharpening, filtering)	LIDAR data processing (filter, ...) preclassify data (satellite and LIDAR data)
Data analysis	perform image classification interpret data using expert knowledge	modelling for data analysis (e.g. hydrological modelling, habitat)	perform change detection combine data for adding context, further information
Validation	assessing accuracy of classification	acquire feedback from users (user validation)	
Final product creation	design look & feel of product (might include templates)	prepare a product description	post-processing of data (format conversion, metadata...) create information products
Manage projects	time keeping contracting, preparing bills etc. -> administer project	manage resources (licenses, data storage, computers...) define project structure	specify detailed project plan re-evaluate the workflow (lessons learnt)
Research and Development	test alternative software products/algorithms	explore/evaluate upcoming developments and trends	develop predefined workflow (e.g. OBIA workflow, preprocessing workflow...)
Professional development	visit conferences	take training (webinars...)	give training
Data management	store data	versioning	share data license data

Analysis ready data
Automation of image analysis
Platform as a service
Information as a service

Fig. 5 Tasks affected by the four identified technological advances in the EO field

5 Scenarios for Future Occupational Profiles

The question that remains to be answered is how the profile of a RS specialist evolves due to the technological advances and the high degree of automation. Figure 6 shows the combination of the two previous figures concerning skills and competencies of RS specialists and the impacted tasks. It can be observed that most tasks that require technical skills are affected by the identified technological advances. Tasks requiring transversal skills remain unaffected by technological developments in the EO domain. Out of 25 tasks that require competencies and experience with EO project, at least 10 are affected by the developments of the domain; some of the tasks might disappear from the profile or become redundant whereas for others online platforms provide the new working environments.

One of the objectives of the EO4GEO project is to derive vocational training and education (VET) programmes that foster user uptake in newly established EO service provider businesses. VET programmes lead to graduates who are equipped with strong technical and transversal skills. Tasks requiring technical skills that can be assigned to less trained workforce are the ones that are likely to be most affected by the technological developments. The question arising from this observation is how VET programmes can be designed to be ‘future ready’.

The second observation that the illustration supports, is, which high level tasks will experience change and therefore might indicate the necessity of retraining RS specialists. PaaS and INFOaaS are strongly promoted and need to be accompanied with training measures. As the development or extension of (existing) training material is part of the EO4GEO project, the change in working environment for accomplishing EO related tasks will have to be addressed.

Figure 7 illustrates, which tasks are expected to be replaced by automated execution or the availability of ARD; these tasks would no longer have to be performed by the RS specialist using today’s tools and procedures. We speak of tasks that are cancelled from the occupational profile of the RS specialist. Cancelling tasks from the profile means that the tasks are automated and require less skills, competencies and resources for their execution; it does not mean that these operations are not required anymore. The question arising from this illustration is how to fill the created gaps in the occupational profile.

6 Discussion and Conclusions

In this contribution we illustrated the impact of developments in the EO sector on a specific profile of a RS specialist. The approach is expert based and constitutes an intuitive forecasting method (Committee on Forecasting Future Disruptive Technologies 2010). The resulting illustrations are subject to uncertainties that are inherent to forecasts. Nevertheless, the illustrations support ongoing discussions in the EO4GEO project on the required training and future occupational profiles in the EO/GI sector.

Remote Sensing Specialist – Change of Workflow

Communication	continuous exchange with users	reporting of progress, results, challenges	prepare and communicate an offer (incl. a quote)	communication within the organization	communication with data providers and other stakeholders
Project conceptualization	analyse needs of users (incl. order info)	define required types of data	define the methodology	Implement (standardised) predefined workflows	create classification schemes (incl. semantics)
Data acquisition	search for data	order archived data (images and vector data)	collect in-field data (GPS, reference data...)	tasking of a custom image (custom parameter settings)	assure data quality
Data preprocessing	radiometric calibration	geometric correction (ortho-rectification, co-registration...)	perform image enhancement (gain-sharpening...)	filtering	preclassify data (satellite and LIDAR data)
Data analysis	perform image classification	interpret data using expert knowledge	modelling for data analysis (e.g. hydrological modelling, habitat)	delineate objects manually (e.g. houses)	perform change detection
Validation	assessing accuracy of classification	acquire feedback from users (user validation)	incorporate feedback		combine data for adding context, further information
Final product creation	design look & feel of product (might include templates)	prepare a product description	post-processing of data (format conversion, metadata...)	create information products (PDF, webmap...)	
Manage projects	time keeping	contracting, preparing bills etc. -> administer project	manage resources (licenses, data storage, computers...)	define project structure	specify detailed project plan
Research and Development	test alternative software products/algorithms	explore/evaluate upcoming developments and trends	develop predefined workflow (e.g. OBIA workflow, preprocessing workflow...)		re-evaluate the workflow (lessons learnt)
Professional development	visit conferences	take training (webinars...)	give training		
Data management	store data	Versioning	share data	license data	

Analysis ready data
Automation of image analysis
Platform as a service
Information as a service

task requiring transversal skills
task requiring technical skills
task requiring competencies

Fig. 6 Combined visualization showing the tasks that are affected by developments as well as the skills or competencies required for the tasks; white boxes indicate tasks that seem unaffected by technological advances

Remote Sensing Specialist – Tasks cancelled from the profile

Communication	continuous exchange with users	reporting of progress, results, challenges	prepare and communicate an offer (incl. a quote)	communication within the organization	communication with data providers and other stakeholders
	analyse needs of users (incl. order info)	define required types of data	define the methodology	Implement (standardised) predefined workflows	create classification schemes (incl. semantics)
Data acquisition	search for data	order archived data (images and vector data)	collect in-field data (GPS, reference data...)	tasking of a custom image (custom parameter settings)	assure data quality
	radiometric calibration	geometric correction (ortho-rectification, co-registration...)	perform image enhancement (pan-sharpening, filtering)	data manipulation, filtering	LIDAR data processing (filter...)
Data analysis	perform image classification	interpret data using expert knowledge	modelling for data analysis (e.g. hydrological modelling, habitat)	delineate objects manually (e.g. houses)	perform change detection
	assessing accuracy of classification	acquire feedback from users (user validation)	incorporate feedback		combine data for adding context, further information
Final product creation	design look & feel of product (moeht include templates)	prepare a product description	post-processing of data (format conversion, metadata...)	create information products (PDF, webmap...)	
	time keeping	contracting, preparing bills etc. -> administer project	manage resources (licenses, data storage, computers...)	define project structure	specify detailed project plan
Research and Development	test alternative software products/algorithms	explore/evaluate upcoming developments and trends	develop predefined workflow (e.g. OBIA workflow, preprocessing workflow...)		
	visit conferences	take training (webinars...)	give training		
Data management	store data	Versioning	share data	license data	

task requiring transversal skills
task requiring technical skills
task requiring competencies

Tasks cancelled due to automation or APD

Fig. 7 Tasks greyed out are expected to be eliminated from the RS specialist profile because of automation or the availability of analysis ready data

Based on the observations outlined in the previous sections, we conclude for the specific profile we analysed:

- The occupational profile of RS specialists will change considerably in the midterm due to developments in the sector.
- The developments may lead to an elimination of some of the tasks that mostly require technical skills from the tasks of an RS specialist. This change suggests that the number of tasks is streamlined and that there is room for filling the gaps with new tasks.
- Workforce with high skill levels and strong competencies in the field will remain essential in the sector.
- Transversal skills related to communication and project management, as well as problem solving, creativity, team work, motivation for retraining etc. as required as part of competencies in the field, will remain important or even increase in importance.

One of the possible scenarios for the future profile of RS specialists is that they focus their work on the development and improvement of platforms, automation routines, preparation of analysis ready data etc. Such profiles will be required (as they are required at the moment as well) and constitute a merged profile of a RS specialist and RS developer. Due to the centralization of efforts, the number of required workforce is expected to be rather small.

Automation and online work environments are expected to reduce the amount of resources required for image analysis and EO service development. There are different scenarios possible to complement the reduced profile:

- One scenario is the expansion of competencies towards application domains and a strengthened exchange with end users.
- Another possible complement is the stronger consideration of entrepreneurial aspects of RS product development.
- A third possible scenario is the focus on new methodologies for the development of value added products, which suggests the exploitation of an even stronger integration of the EO and GI sectors.

The listed possibilities to complement tasks of current RS specialist all point to a more interdisciplinary profile. Workforce with interdisciplinary competencies and solid transversal skills can be expected to steer the uptake of the EO sector in the future.

Regarding the development of new VET programmes for the sector, the present analysis cannot foresee which tasks might be assigned to the workforce with the respective training. This depends on the particular scenario that evolves and the tasks at suitable skill and competency levels that will then be added to the profile.

Partners of the EO4GEO project currently work on the analysis of further profiles like EO application developer, business analyst, data provider or senior project manager according to the methodology presented in this contribution. The expansion of the analysis to additional profiles can provide deeper insights into the expected transformation of the EO/GI sector. The results of the profile analysis are input for the

development of an overarching sector skill strategy and the development of training material for the future EO/GI workforce as foreseen in the project.

Acknowledgements We kindly acknowledge the participation of the following remote sensing specialists in the DACUM workshop and their support of this work: Sebastian d'Oleire-Oltmanns, Kerstin Kulessa, Gina Schwendemann and Thomas Strasser. We also acknowledge the input received from Peter Zeil during the development of the presented material. Comments by anonymous reviewers improved the content of the paper. This work has partially been supported by the Erasmus + Sector Skills Alliance project EO4GEO.

References

- Aguilar Moreno E, Hofer B, Lang S (2018) D 1.2—workshop on demand for space/geospatial education and training. Deliverable of ERASMUS + Project EO4GEO—towards an innovative strategy for skills development and capacity building in the space geoinformation sector supporting Copernicus user uptake. <http://www.eo4geo.eu/publications/>
- Cedefop (2013) Quantifying skill needs in Europe. Occupational skills profiles: methodology and application. European Centre for the Development of Vocational Training, Luxembourg
- Cedefop (2014) Terminology of European education and training policy: a selection of 130 terms, 2nd edn. Publications Office, Luxembourg
- Chen J, Dowman I, Li S, Li Z, Madden M, Mills J, Paparoditis N, Rottensteiner F, Sester M, Toth C, Trinder J, Heipke C (2016) Information from imagery: ISPRS scientific vision and research agenda. *ISPRS J Photogramm Remote Sens* 115:3–21
- Collum J (1999) Analyse von Berufen mit dem DACUM-Prozess. Panorama - Fachinformation für Berufsbildung, Berufsberatung und Arbeitsmarkt, p 1
- Committee on Forecasting Future Disruptive Technologies (2010) Persistent forecasting of disruptive technologies. The National Academies Press
- European Commission (2016) Copernicus. Market report. Publications Office of the European Union, Luxembourg
- Guo H (2017) Big Earth data: a new frontier in Earth and information sciences. *Big Earth Data* 1(1–2):4–20
- Hofer B, d'Oleire Oltmanns S, Ferber N, Albrecht F, Lang S (2018) D 1.3—demand for space/geospatial education and training and priority occupational profiles. Deliverable of ERASMUS + Project EO4GEO—towards an innovative strategy for skills development and capacity building in the space geoinformation sector supporting Copernicus user uptake. <http://www.eo4geo.eu/publications/>
- Johnson J (2010) What GIS technicians do: a synthesis of DACUM job analyses. *J Urban Reg Inf Syst Assoc* 22(2):31–40
- Kunz R (2015) Wissen und Handeln in Schlüsselsituationen der Sozialen Arbeit - Empirische und Theoretische Grundlegung eines neuen kausistischen Ansatzes. Unpublished Ph.D., Universität Basel, Basel
- Lang S, Baraldi A, Tiede D, Hay G, Blaschke T (2018) Towards a GEOBIA 2.0 manifesto? Achievements and open challenges in information & knowledge extraction from big Earth data. In: Paper presented to the GEOBIA 2018—from pixels to ecosystems and global sustainability, Montpellier, France
- Li S, Dragicevic S, Castro FA, Sester M, Winter S, Coltekin A, Pettit C, Jiang B, Haworth J, Stein A, Cheng T (2016) Geospatial big data handling theory and methods: a review and research challenges. *ISPRS J Photogramm Remote Sens* 115:119–133
- Norton RE (1997) DACUM handbook, 2 edn. Center on Education and Training for Employment, College of Education, The Ohio State University, Columbus, Ohio

- O'Sullivan C, Wise N, Mathieu P-P (2018) The changing landscape of geospatial information markets. In: Mathieu P-P, Aubrecht C (eds) *Earth observation open science and innovation*. Springer International Publishing, Cham, pp 3–23
- Sudmanns M, Lang S, Tiede D (2018) Big earth data: from data to information. *GI_Forum* 2018(1):184–193
- Tippelt R, Edelman D (2007) DACUM (developing a curriculum). In: Erpenbeck J, von Rosenstiel L (eds.) *Handbuch Kompetenzmessung. Erkennen, Verstehen und Bewerten von Kompetenzen in der betrieblichen, pädagogischen und psychologischen Praxis*. Schäffer-Poesche, Stuttgart, pp 737–757
- Wallentin G, Hofer B, Traun C (2015) Assessment of workforce demands to shape GIS&T education. *Trans GIS* 19(3):439–454