

13th International Conference on Geographic Information Science

GIScience 2025, August 26–29, 2025, Christchurch, New Zealand

Edited by

Katarzyna Sila-Nowicka

Antoni Moore

David O'Sullivan

Benjamin Adams

Mark Gahegan



Editors

Katarzyna Sila-Nowicka 

University of Auckland, New Zealand
katarzyna.sila-nowicka@auckland.ac.nz

Antoni Moore 

University of Otago, New Zealand
tony.moore@otago.ac.nz

David O'Sullivan 

University of Auckland, New Zealand
osullivan512@gmail.com

Benjamin Adams 

University of Canterbury, Christchurch, New Zealand
benjamin.adams@canterbury.ac.nz

Mark Gahegan 

University of Auckland, New Zealand
m.gahegan@auckland.ac.nz

ACM Classification 2012

Information systems → Spatial-temporal systems; Computing methodologies → Machine learning;
Computing methodologies → Modeling and simulation

ISBN 978-3-95977-378-2

Published online and open access by

Schloss Dagstuhl – Leibniz-Zentrum für Informatik GmbH, Dagstuhl Publishing, Saarbrücken/Wadern,
Germany. Online available at <https://www.dagstuhl.de/dagpub/978-3-95977-378-2>.

Publication date

August, 2025

Bibliographic information published by the Deutsche Nationalbibliothek

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

License

This work is licensed under a Creative Commons Attribution 4.0 International license (CC-BY 4.0):
<https://creativecommons.org/licenses/by/4.0/legalcode>.



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

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

The copyright is retained by the corresponding authors.

Digital Object Identifier: 10.4230/LIPICS.GIScience.2025.0

ISBN 978-3-95977-378-2

ISSN 1868-8969

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

LIPIcs – Leibniz International Proceedings in Informatics

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

Editorial Board

- Christel Baier (TU Dresden, DE)
- Roberto Di Cosmo (Inria and Université Paris Cité, FR)
- Faith Ellen (University of Toronto, CA)
- Javier Esparza (TU München, DE)
- Holger Hermanns (Universität des Saarlandes, Saarbrücken, DE and Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Wadern, DE)
- Daniel Král' (Leipzig University, DE and Max Planck Institute for Mathematics in the Sciences, Leipzig, DE)
- Sławomir Lasota (University of Warsaw, PL)
- Meena Mahajan (Institute of Mathematical Sciences, Chennai, IN – Chair)
- Chih-Hao Luke Ong (Nanyang Technological University, SG)
- Eva Rotenberg (Technical University of Denmark, Lyngby, DK)
- Pierre Senellart (ENS, Université PSL, Paris, France)
- Alexandra Silva (Cornell University, Ithaca, US)

ISSN 1868-8969

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

■ Contents

Preface

<i>Katarzyna Sila-Nowicka, Antoni Moore, David O'Sullivan, Benjamin Adams, and Mark Gahegan</i>	0:vii
---	-------

Reviewers

.....	0:xiii
-------	--------

List of Authors

.....	0:xiii
-------	--------

Regular Papers

Leveraging Open-Source Satellite-Derived Building Footprints for Height Inference <i>Clinton Stipek, Taylor Hauser, Justin Epting, Jessica Moehl, and Daniel Adams</i> ..	1:1–1:20
CityJSON Management Using Multi-Model Graph Database to Support 3D Urban Data Management <i>Muhammad Syafiq, Suhaibah Azri, and Uznir Ujang</i>	2:1–2:15
Enriching Location Representation with Detailed Semantic Information <i>Junyuan Liu, Xinglei Wang, and Tao Cheng</i>	3:1–3:15
Precomputed Topological Relations for Integrated Geospatial Analysis Across Knowledge Graphs <i>Katrina Schweikert, David K. Kedrowski, Shirley Stephen, and Torsten Hahmann</i>	4:1–4:22
Analysis of Points of Interests Recommended for Leisure Walk Descriptions <i>Ehsan Hamzei, Thi Minh Hoai Bui, Martin Tomko, and Stephan Winter</i>	5:1–5:16
MODAP: A Multi-City Open Data & Analytics Platform for Micromobility Research <i>Grant McKenzie</i>	6:1–6:14
A Modularity-Driven Framework for Unraveling Congestion Centers with Enhanced Spatial-Semantic Features <i>Weihua Huan, Xintao Liu, and Wei Huang</i>	7:1–7:11
BERT4Traj: Transformer-Based Trajectory Reconstruction for Sparse Mobility Data <i>Hao Yang, Angela Yao, Christopher C. Whalen, and Gengchen Mai</i>	8:1–8:9
Identifying Resilient Communities in Road Networks: A Path-Based Embedding Approach <i>Christopher Wagner, Somayeh Dodge, and Danial Alizadeh</i>	9:1–9:10
Geovicla: Automated Classification of Interactive Web-Based Geovisualizations <i>Phil Hüffer, Auriol Degbelo, and Benjamin Risse</i>	10:1–10:12
Georeferencing Historical Maps at Scale <i>Rere-No-A-Rangi Pope and Marcus Freen</i>	11:1–11:11

13th International Conference on Geographic Information Science (GIScience 2025).

Editors: Katarzyna Sila-Nowicka, Antoni Moore, David O'Sullivan, Benjamin Adams, and Mark Gahegan



Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Large Multi-Modal Model Cartographic Map Comprehension for Textual Locality Georeferencing	
<i>Kalana Wijegunaratna, Kristin Stock, and Christopher B. Jones</i>	12:1–12:19
Assessing Map Reproducibility with Visual Question-Answering: An Empirical Evaluation	
<i>Eftychia Koukouraki, Auriol Degbelo, and Christian Kray</i>	13:1–13:12
Guiding Geospatial Analysis Processes in Dealing with Modifiable Areal Unit Problems	
<i>Guoray Cai and Yue Hao</i>	14:1–14:18
Accommodating Space-Time Scaling Issues in GAM-Based Varying Coefficient Models	
<i>Alexis Comber, Paul Harris, and Chris Brunsdon</i>	15:1–15:9
What, When, and Where Do You Mean? Detecting Spatio-Temporal Concept Drift in Scientific Texts	
<i>Meilin Shi, Krzysztof Janowicz, Zilong Liu, Mina Karimi, Ivan Majic, and Alexandra Fortacz</i>	16:1–16:18
The Inherent Structure of Experiments as a Constraint to Spatial Analysis and Modeling	
<i>Simon Scheider and Judith A. Verstegen</i>	17:1–17:17
U-Prithvi: Integrating a Foundation Model and U-Net for Enhanced Flood Inundation Mapping	
<i>Vit Kostejn, Yamil Essus, Jenna Abrahamson, and Ranga Raju Vatsavai</i>	18:1–18:17
Search Space Reduction Using Species Distribution Modeling with Simulated Pollen Signatures	
<i>Haoyu Wang, Jennifer A. Miller, and Shalene Jha</i>	19:1–19:6

■ Preface

This volume contains the full paper proceedings of the 13th International Conference on Geographic Information Science (GIScience 2025), held at University of Canterbury, Christchurch, with strong support from the GIScience community at the University of Auckland, University of Otago, and Victoria University Wellington, 26th to 29th August 2025.

This is the first time the GIScience conference has been held in New Zealand and the second time in the southern hemisphere. We received 47 submissions, each reviewed by three to five members of the international programme committee, with 19 full papers being accepted. In addition to these papers, 67 abstract papers, 6 demos, and 76 posters were accepted for presentation at the conference.

The accepted papers represent a wide range of topics across GIScience, including analysis and modelling on urban data, mobility and transportation, semantic enrichment, automated map georeferencing and classification, statistical modelling in space and time, and environmental modelling applications. Separate to the main conference programme was a group of pre-conference workshops, eight research seminar, tutorial or other participatory events that again covered a breadth of GIScience topics.

The entire GIScience 2025 team would like to express their gratitude to all the authors, reviewers, and workshop organisers, and everyone else involved in the conference, including the sponsors and keynotes.

Platinum Sponsors



Gold and Silver Sponsors



Taylor & Francis Group
an informa business



Toitū Te Whenua
Land Information
New Zealand

13th International Conference on Geographic Information Science (GIScience 2025).

Editors: Katarzyna Sila-Nowicka, Antoni Moore, David O'Sullivan, Benjamin Adams, and Mark Gahegan



Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Reviewers

- Benjamin Adams, University of Canterbury
- Clio Andris, Georgia Institute of Technology
- Jagannath Aryal, The University of Melbourne
- Crystal Bae, University of Chicago
- Joana Barros, Birkbeck, University of London
- Mary-Kate Beard-Tisdale, University of Maine
- Michela Bertolotto, University College Dublin
- Filip Biljecki, National University of Singapore
- Lars Bodum, Aalborg University
- Vanessa Brum-Bastos, University of Canterbury
- Pedro Cabral, Universidade Nova de Lisboa
- Christophe Claramunt, Naval Academy Research Institute
- Eliseo Clementini, University of Aquila
- Lex Comber, University of Leeds
- Clodoveu Davis, Federal University of Minas Gerais
- Urska Demsar, University of St. Andrews
- Mairead de Róiste, Victoria University of Wellington
- Stef De Sabbata, University of Leicester
- Somayeh Dodge, University of California, Santa Barbara
- Suzana Dragicevic, Simon Fraser University
- Ekaterina Egorova, University of Twente
- Sara Fabrikant, University of Zurich
- Mark Gahegan, University of Auckland
- Song Gao, University of Wisconsin - Madison
- Michael T. Gastner, Singapore Institute of Technology
- Ioannis Giannopoulos, Technische Universität Wien
- Amy Griffin, Royal Melbourne Institute of Technology
- Torsten Hahmann, University of Maine
- Serene Ho, University of Melbourne
- Hartwig Hochmair, University of Florida
- Bernhard Höfle, Ruprecht-Karls-Universität Heidelberg
- Yingjie Hu, State University of New York at Buffalo
- Carolynne Hultquist, University of Canterbury
- Piotr Jankowski, San Diego State University
- Krzysztof Janowicz, Universität Vienna
- Christopher B. Jones, Cardiff University
- Tomi Kauppinen, Aalto University
- Pyry Kettunen, National Land Survey of Finland
- Peter Kiefer, ETH Zurich
- Carsten Keßler, Hochschule Bochum – Bochum University of Applied Sciences
- Minh Kieu, University of Auckland
- Dimitris Kotzinos, CY Cergy Paris University
- Christian Kray, University of Münster
- Shawn Laffan, University of New South Wales
- Arika Ligmann-Zielinska, Michigan State University

13th International Conference on Geographic Information Science (GIScience 2025).

Editors: Katarzyna Sila-Nowicka, Antoni Moore, David O'Sullivan, Benjamin Adams, and Mark Gahegan




Leibniz International Proceedings in Informatics


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


- Samsung Lim, University of New South Wales
- Yan Liu, The University of Queensland
- Binbin Lu, Wuhan University
- Gengchen Mai, University of Texas at Austin
- Nick Malleson, University of Leeds
- Bruno Martins, Instituto Superior Técnico
- Grant McKenzie, McGill University
- Liqu Meng, Technische Universität München
- Harvey Miller, Ohio State University
- Jennifer Miller, University of Texas at Austin
- Franz-Benjamin Mocnik, Paris Lodron Universität Salzburg
- Daniel R. Montello, University of California, Santa Barbara
- Antoni Moore, University of Otago
- Mir Abolfazl Mostafavi, Université Laval
- Alan Murray, University of California, Santa Barbara
- Atsushi Nara, San Diego State University
- Javier Nogueras Iso, University of Zaragoza
- David O'Sullivan, University of Auckland
- Volker Paelke, Dominica State College
- Manon Prédhumeau, University of Leeds
- Ross Purves, University of Zurich
- Tumasch Reichenbacher, University of Zurich
- Claus Rinner, Toronto Metropolitan University
- Anthony C. Robinson, Pennsylvania State University
- Armanda Rodrigues, Universidade Nova de Lisboa
- Simon Scheider, Utrecht University
- Oliver Schmitz, Utrecht University
- Johannes Schöning, University of St. Gallen
- Johannes Scholz, Paris Lodron University Salzburg
- Angela Schwering, Westfälische Wilhelms-Universität Münster
- Raja Sengupta, McGill University
- Monika Sester, Universität Hannover
- Shih-Lung Shaw, University of Tennessee, Knoxville
- Hyesop Shin, University of Auckland
- Takeshi Shirabe, KTH Royal Institute of Technology
- Gaurav Sinha, Ohio University
- Katarzyna Sila-Nowicka, University of Auckland
- Yongze Song, Curtin University
- Kathleen Stewart, University of Maryland, College Park
- Martin Swobodzinski, Portland State University
- Sabine Timpf, University of Augsburg
- Martin Tomko, University of Melbourne
- Ming-Hsiang Tsou, San Diego State University
- Nico Van de Weghe, Universiteit Gent
- Marc J. van Kreveld, Utrecht University
- Judith Verstegen, Utrecht University
- May Yuan, University of Texas at Dallas
- John Wilson, University of Southern California


- Stephan Winter, University of Melbourne
- Ningchuan Xiao, The Ohio State University
- Jing Yao, University of Glasgow
- Qunshan Zhao, University of Glasgow
- Rui Zhu, University of Bristol
- Sisi Zlatanova, University of New South Wales


■ List of Authors

Jenna Abrahamson  (18)
Center for Geospatial Analytics, North Carolina
State University, Raleigh, NC, USA


Daniel Adams  (1)
Geospatial Science and Human Security Division,
Oak Ridge National Laboratory, TN, USA

Danial Alizadeh  (9)
Department of Geography, University of
California, Santa Barbara, CA, USA


Suhaibah Azri  (2)
3D GIS Research Lab, Universiti Teknologi
Malaysia, Johor Bahru, Malaysia


Chris Brunsdon  (15)
National Centre for Geomcomputation, National
University of Ireland, Maynooth, Ireland


Thi Minh Hoai Bui  (5)
University of Melbourne, Parkville, Australia

Guoray Cai  (14)
College of Information Sciences and Technology,
The Pennsylvania State University, University
Park, PA, USA


Tao Cheng  (3)
SpaceTimeLab, University College London, UK


Alexis Comber  (15)
School of Geography, University of Leeds, UK


Auriol Degbelo  (10, 13)
Chair of Geoinformatics, TU Dresden, Germany


Somayeh Dodge  (9)
Department of Geography, University of
California, Santa Barbara, CA, USA

Justin Epting  (1)
Bechtel National, Inc., Ogden, UT, USA


Yamil Essus  (18)
Industrial and Systems Engineering Department,
North Carolina State University, Raleigh, NC,
USA


Alexandra Fortacz  (16)
Department of Geography and Regional
Research, University of Vienna, Austria


Marcus Frean  (11)
Victoria University of Wellington, Aotearoa,
New Zealand


Torsten Hahmann  (4)
School of Computing and Information Science,
University of Maine, Orono, ME, USA


Ehsan Hamzei  (5)
University of Melbourne, Parkville, Australia


Yue Hao  (14)
College of Information Sciences and Technology,
The Pennsylvania State University,
University Park, PA, USA

Paul Harris  (15)
Sustainable Agriculture Sciences,
Rothamsted Research, Harpenden, UK


Taylor Hauser  (1)
Geospatial Science and Human Security Division,
Oak Ridge National Laboratory, TN, USA

Weihua Huan  (7)
College of Surveying and Geo-informatics,
Tongji University, Shanghai, China;
Department of Land Surveying and
Geo-Informatics, The Hong Kong Polytechnic
University, Kowloon, Hong Kong


Wei Huang  (7)
College of Surveying and Geo-informatics,
Tongji University, Shanghai, China;
Department of Civil Engineering, Toronto
Metropolitan University, Canada;
Urban Mobility Institute, Tongji University,
Shanghai, China

Phil Hüffer  (10)
Institute for Geoinformatics,
University of Münster, Germany


Krzysztof Janowicz (16)
Department of Geography and Regional
Research, University of Vienna, Austria


Shalene Jha  (19)
Department of Integrative Biology,
University of Texas at Austin, TX, USA

Christopher B. Jones  (12)
School of Computer Science and Informatics,
Cardiff University, UK

Mina Karimi  (16)
Department of Geography and Regional
Research, University of Vienna, Austria

- David K. Kedrowski  (4)
School of Computing and Information Science,
University of Maine, Orono, ME, USA
- Vit Kostejn  (18)
Charles University, Prague, Czech Republic
- Eftychia Koukouraki  (13)
Institute for Geoinformatics,
University of Münster, Germany
- Christian Kray  (13)
Institute for Geoinformatics,
University of Münster, Germany
- Junyuan Liu  (3)
SpaceTimeLab, University College London, UK
- Xintao Liu  (7)
Department of Land Surveying and
Geo-Informatics, The Hong Kong Polytechnic
University, Kowloon, Hong Kong
- Zilong Liu  (16)
Department of Geography and Regional
Research, University of Vienna, Austria
- Gengchen Mai (8)
Department of Geography and the Environment,
University of Texas at Austin, TX, USA
- Ivan Majic  (16)
Department of Geography and Regional
Research, University of Vienna, Austria
- Grant McKenzie  (6)
Platial Analysis Lab, McGill University,
Montréal, Canada
- Jennifer A. Miller  (19)
Department of Geography and the Environment,
University of Texas at Austin, TX, USA
- Jessica Moehl  (1)
Geospatial Science and Human Security Division,
Oak Ridge National Laboratory, TN, USA
- Rere-No-A-Rangi Pope  (11)
Victoria University of Wellington,
Aotearoa, New Zealand
- Benjamin Risse  (10)
Institute for Geoinformatics, University of
Münster, Germany
- Simon Scheider  (17)
Department of Human Geography and Spatial
Planning, Utrecht University, The Netherlands
- Katrina Schweikert  (4)
School of Computing and Information Science,
University of Maine, Orono, ME, USA
- Meilin Shi  (16)
Department of Geography and Regional
Research, University of Vienna, Austria
- Shirly Stephen  (4)
NCEAS, Department of Geography, University
of California, Santa Barbara, CA, USA;
School of Computing and Information Science,
University of Maine, Orono, ME, USA
- Clinton Stipek  (1)
Geospatial Science and Human Security Division,
Oak Ridge National Laboratory, TN, USA
- Kristin Stock  (12)
School of Mathematical and Computational
Sciences, Massey University, Auckland, New
Zealand
- Muhammad Syafiq  (2)
3D GIS Research Lab, Universiti Teknologi
Malaysia, Johor Bahru, Malaysia
- Martin Tomko  (5)
University of Melbourne, Parkville, Australia
- Uznir Ujang  (2)
3D GIS Research Lab, Universiti Teknologi
Malaysia, Johor Bahru, Malaysia
- Ranga Raju Vatsavai  (18)
Computer Science Department, North Carolina
State University, Raleigh, NC, USA
- Judith A. Verstegen  (17)
Department of Human Geography and Spatial
Planning, Utrecht University, The Netherlands
- Christopher Wagner  (9)
Department of Statistics and Applied
Probability, University of California, Santa
Barbara, CA, USA
- Haoyu Wang  (19)
Department of Geography and the Environment,
University of Texas at Austin, TX, USA
- Xinglei Wang  (3)
SpaceTimeLab, University College London, UK
- Christopher C. Whalen  (8)
College of Public Health, University of Georgia,
Athens, GA, USA

Kalana Wijegunaratna  (12)
School of Mathematical and Computational
Sciences, Massey University, Auckland, New
Zealand

Stephan Winter  (5)
University of Melbourne, Parkville, Australia

Hao Yang (8)
Department of Geography, University of Georgia,
Athens, GA, USA

Angela Yao (8)
Department of Geography, University of Georgia,
Athens, GA, USA

Leveraging Open-Source Satellite-Derived Building Footprints for Height Inference

Clinton Stipek¹ ✉ 

Geospatial Science and Human Security Division, Oak Ridge National Laboratory, TN, USA

Taylor Hauser ✉ 

Geospatial Science and Human Security Division, Oak Ridge National Laboratory, TN, USA

Justin Epting ✉ 

Bechtel National, Inc., Ogden, UT, USA

Jessica Moehl ✉ 

Geospatial Science and Human Security Division, Oak Ridge National Laboratory, TN, USA

Daniel Adams ✉ 

Geospatial Science and Human Security Division, Oak Ridge National Laboratory, TN, USA

Abstract

At a global scale, cities are growing and characterizing the built environment is essential for deeper understanding of human population patterns, urban development, energy usage, climate change impacts, among others. Buildings are a key component of the built environment and significant progress has been made in recent years to scale building footprint extractions from satellite datum and other remotely sensed products. Billions of building footprints have recently been released by companies such as Microsoft and Google at a global scale. However, research has shown that depending on the methods leveraged to produce a footprint dataset, discrepancies can arise in both the number and shape of footprints produced. Therefore, each footprint dataset should be examined and used on a case-by-case study. In this work, we find through two experiments on Oak Ridge National Laboratory and Microsoft footprints within the same geographic extent that our approach of inferring height from footprint morphology features is source agnostic. Regardless of the differences associated with the methods used to produce a building footprint dataset, our approach of inferring height was able to overcome these discrepancies between the products and generalize, as evidenced by 98% of our results being within 3m of the ground-truthed height. This signifies that our approach can be applied to the billions of open-source footprints which are freely available to infer height, a key building metric. This work impacts the broader domain of urban science in which building height is a key, and limiting factor.

2012 ACM Subject Classification Computing methodologies → Machine learning; Computing methodologies → Classification and regression trees; Computing methodologies → Neural networks; Applied computing

Keywords and phrases Building Height, Big Data, Machine Learning

Digital Object Identifier 10.4230/LIPIcs.GIScience.2025.1

Acknowledgements Notice: This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (<http://energy.gov/downloads/doe-public-access-plan>).

¹ Corresponding author



© Clinton Stipek, Taylor Hauser, Justin Epting, Jessica Moehl, and Daniel Adams; licensed under Creative Commons License CC-BY 4.0

13th International Conference on Geographic Information Science (GIScience 2025).

Editors: Katarzyna Sila-Nowicka, Antoni Moore, David O'Sullivan, Benjamin Adams, and Mark Gahegan; Article No. 1; pp. 1:1–1:20



Leibniz International Proceedings in Informatics
LIPIcs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

Populations are increasing at a global scale, and it is estimated that by 2050, 68% of the global population will live in urban environments [17, 20]. Buildings are a key component of the urban environment and their footprints have been used across a myriad of subjects such as population density estimation [31, 32], building energy usage [12, 39], disaster management [28], building type [1], building height [6, 25, 37] and urban heat islands (UHI's) [10]. Being able to characterize the built environment is imperative to address these issues and information on building footprints, building height and urban morphology is critical in these efforts.

Fortunately, over the past decade, there has been a dramatic increase in the amount of open-source building footprint datum available from organizations such as Microsoft², Google³, and Oak Ridge National Laboratory (ORNL)⁴. Between these products, there are over 3 billion footprints available for use. However, each of the aforementioned products is generated via differing methods for the pixel extraction/segmentation to identify buildings and the regularization process of the identified footprints. Furthermore, differences in imagery sources and resolution as well as environmental factors such as shadows or sun angle can also influence the footprint extraction and regularization process [13].

The processing workflow from Microsoft is described in their documentation as first leveraging a deep neural network (DNN) to identify buildings from aerial imagery and then converting the identified pixels into polygons representing building footprints⁵. The best available information on Google's footprint generation is from a paper in 2023 by Sirko et al. [35]. In their report, the authors describe utilizing a U-Net model, a common approach for segmenting satellite datum [2, 29, 38]. Once extracted, the building footprints are then processed through a contouring algorithm that realigns groups of adjacent polygons to regularize the building footprints [35]. The authors also provide a caveat that newer versions (v2 and v3) of the Google Open Buildings dataset underwent further improvements that are not documented. Of the three datasets, ORNL provides the highest level of detail in how building footprints were both extracted and regularized from satellite datum as they leverage a deep convolutional neural network (CNN) framework for pixel extraction and the ArcGIS proprietary building footprint regularization module⁶ [41, 42]. However, the ORNL footprint dataset is only available publicly in the United States (U.S.), so it lacks the volume and spatial scale of building footprints that Microsoft and Google provide. While Microsoft, Google, and ORNL each utilize a deep learning framework to identify, delineate, and regularize the footprints, there are proven differences associated with the footprints provided by each entity [8, 14].

Chamberlain et al. found substantial differences between footprint patterns displayed by Microsoft and Google when comparing the products at a grid scale in Africa [8]. The authors noted that consideration is needed by users regarding the suitability of the specific building footprint dataset for its intended application. For example, in urban areas, Microsoft seemed to have better coverage in relation to the number of matching footprints, but this pattern was not universal. Also in Africa, Gonzales found patterns of irregularity when comparing Google

² <https://www.microsoft.com/en-us/maps/bing-maps/building-footprints>

³ <https://sites.research.google/open-buildings/>

⁴ <https://gis-fema.hub.arcgis.com/pages/usa-structures>

⁵ <https://github.com/microsoft/GlobalMLBuildingFootprints>

⁶ <https://pro.arcgis.com/en/pro-app/latest/tool-reference/3d-analyst/regularize-building-footprint.htm>

and Microsoft, but at a building-by-building level [14]. When investigating urban areas, Microsoft tended to generate larger footprints which may encapsulate multiple buildings while Google seemed to have more, smaller buildings. When rural areas were investigated, the building counts were relatively similar, showing little discrepancy. Both at scale and at a building-by-building level, care must be taken when leveraging footprint datasets [8, 14].

Recently, research has shown that building height is obtainable based on building footprint information alone [37]. The authors showcased a novel method to infer height at a high accuracy using only information derived from an individual buildings footprint. Furthermore, they did so based on footprints extracted from both lidar and satellite datum. The ability to infer height from footprints extracted from satellite datum allows for this approach to generate building height maps at large scale. However, the authors only demonstrated their approach on building footprints developed by ORNL and the model inference may produce irregular results when exposed to a different footprint source. For example, in Stipek et al. [37] they discuss that the features with the highest impact on inferring building height were contextual (number of neighbors) and engineered (complexity of footprint shape). It has been proven that at both a grid [8] and building-by-building level [14], Microsoft and Google have differing shapes and sizes which would affect the contextual and engineered metrics generated at the building level. Therefore, it would be imprudent to assume the approach proposed by Stipek et al. [37] can be applied to other footprint datasets without further testing.

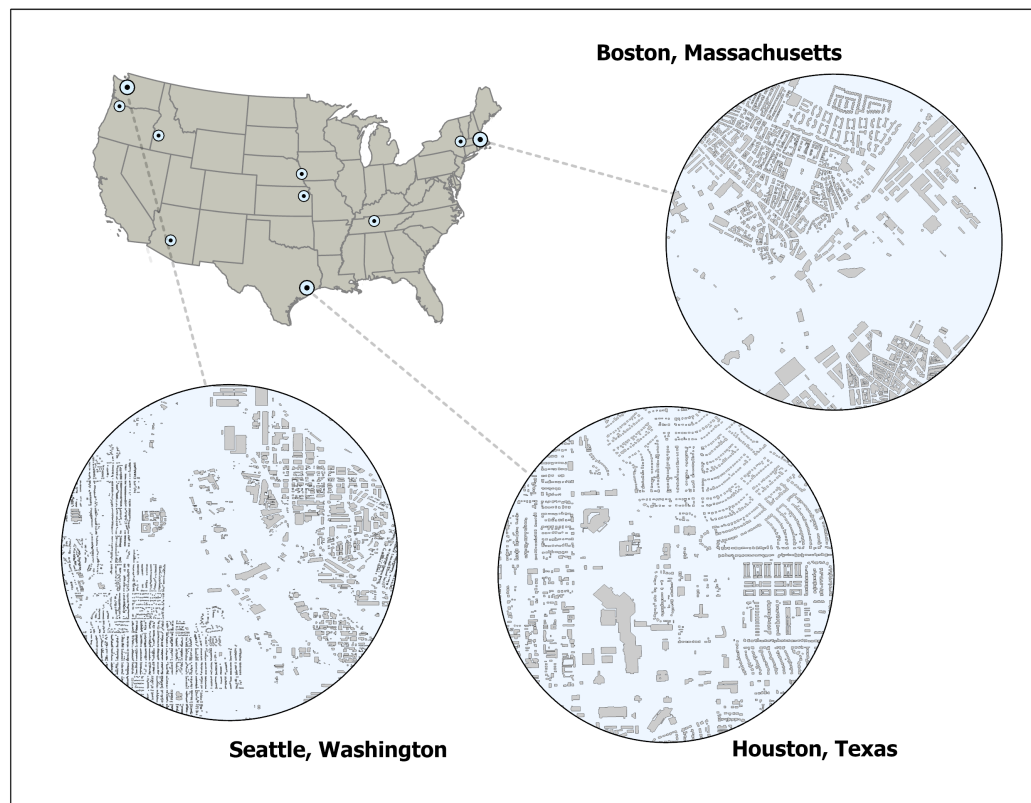
In this paper we demonstrate that it is possible to infer building height at a building-by-building level, agnostic to footprint source. We prove this by comparing the inferred heights from two distinct products, ORNL and Microsoft across 10 cities in the U.S., with figure 1 showcasing the spatial extent of our research. We show that regardless of differences associated with the extraction, regularization of the footprints and other factors, such as imagery date or environmental factors, our approach of inferring height from footprints can be applied to ORNL and Microsoft footprints. This signifies that it is possible to leverage the 1.2 billion footprints which Microsoft has made openly available to infer height at a global scale. Google footprints are not currently available in the U.S. and this research only focused on ORNL and Microsoft footprints.

2 Related Works

While the authors acknowledge a deep field of literature in relation to leveraging deep learning on satellite datum, we would like to bring attention to works which relate to extracting building information. Secondly, we focus on studies that have inferred height from features derived from building footprints.

2.1 Footprint Extraction and Regularization

There have been various methods to segment and regularize footprints derived from high-resolution satellite datum [4, 9, 23, 26, 30, 34, 35, 36, 40, 42, 43]. Shi et al. leverage a large-scale deep learning mapping framework using Google Earth images to map 280 million building footprints in east Asia [34]. They note in their work that existing building extraction models primarily utilize supervised deep learning methods which lack generalization due to differences in building morphologies. For example, buildings in east Asia are more compact and display more diverse patterns as compared to buildings within the U.S. or Europe. The authors further discuss the issues associated with the regularization of the identified buildings, stating that building footprints differ based on the methods leveraged. To address



■ **Figure 1** Map indicating the location of the ten cities used in this study, as well as views of building footprints in Seattle, Houston, and Boston. Note the different patterns within the built environment in each city, a visual representation of the challenges associated with modeling building height.

for this, after the footprints are extracted from the satellite datum, they deploy a stable boundary optimization algorithm which uses a generative adversarial learning network (GAN) to enhance the semantic features of buildings. To regularize the footprints, they used a post-processing method proposed by Gribov [16].

Sirko et al., in developing Google's Open Buildings dataset, leveraged a U-Net architecture, a deep learning encoder-decoder model for semantic segmentation for pixel identification from imagery [33, 35]. This approach classifies each pixel of an image as either a building or a non-building. They tested this approach using a training set of 99,902 satellite images across the African continent and note that two of the more complex issues they faced were smaller buildings, and buildings in densely populated areas. To address for this, they taught their model to predict at least one pixel gap between the buildings, which they accomplished by employing a morphological erosion operation with a kernel size of 3x3 pixels during pre-processing. Once footprints were identified and pre-processed, they then deployed a contouring algorithm to produce angular shapes and realign groups of nearby polygons.

For the development of the ORNL building footprint dataset, Yang et al. developed a CNN framework to extract pixels which represented structures [41, 42]. Furthermore, the authors also incorporated custom designed signed-distance labels which aided in improving the building outline extraction which was especially helpful in core urban areas where there are high densities of buildings within a small area. Once footprints were identified, they

leveraged the ArcGIS building footprint regularization module. Using this approach, the authors provided a simple and effective method that successfully produced a building footprint map of the U.S.

2.2 Building Height - Machine Learning Approach from Morphology Features

In 2017, Biljecki et al. leveraged a random forest model to infer building height from footprint and ancillary information, such as number of floors [6]. The authors tested their approach on 200,000 buildings in the Netherlands and found that it is possible to infer height using a tree-based approach. However, some of the features used, such as number of floors, are a proxy for height and this metric is not available at scale, thus limiting the scalability of their approach. Furthermore, their analysis was done on footprints extracted from lidar datum, thus further hindering them to areas in which lidar footprints are available.

Milojevic-Dupont built upon this work and leveraged a gradient boosting algorithm, XGBoost, to infer height for buildings in Europe (Germany, Netherlands, France, Italy) [25]. They expanded the morphology features derived from building footprints compared to Biljecki [6], and also used ancillary datasets, such as road networks, with a total of 152 features used in their modeling approach. However, similar to the approach by Biljecki, they utilized footprints derived from lidar datum, thus suffering from the same constraint of limited scalability at a continental or global scale.

Stipek et al. expanded upon the work done by the previous authors and inferred height from building footprints without the use of ancillary information [37]. The authors leveraged morphology features generated from individual buildings and successfully inferred height on both lidar-derived and satellite-derived building footprints. However, they only showcased their ability to infer height on ORNL footprints, thus limiting their approach to that singular dataset.

3 Methods

Here, we discuss the methodology used for our research which aims to address if it is possible to infer height based on footprints derived from satellite datum which have been identified and regularized using varying methods. For this work we leverage 3.09 million building records across the U.S. (Table 1). We selected 10 cities within the U.S. that had satellite derived footprints from ORNL and Microsoft which overlapped with lidar derived footprints, which had a height associated with the footprints (Figure 1).

3.1 Footprint Datasets

3.1.1 ORNL Footprints

This dataset contains footprints for the cities of Albany, Boise, Boston, Houston, Nashville, Omaha, Phoenix, Portland, Seattle, and Topeka within the U.S. The footprints are derived from satellite datum based on the approach in Yang et al. [42]. Please note that the current version of footprints within the USA Structures dataset are lidar generated footprints after replacement for the reasons described by Yang et al. [41]. However, we chose to use the earlier version of the satellite derived footprints for fair comparison.

3.1.2 Microsoft Footprints

The Microsoft dataset provides over 1 billion footprints spanning multiple continents⁷. These building footprints were developed from a deep learning model which extracted pixel information from satellite datum. The pixels, once identified as a building, then underwent a thorough cleaning and regularization process. Microsoft has multiple releases for their building footprint dataset and we leverage the footprints Microsoft released on 26/04/2023 for the 10 cities within the U.S. (Table S1).

3.1.3 Lidar Footprints

The lidar footprints leveraged in this research are publicly available as part of the USA Structures dataset at the FEMA portal⁸ [41].

3.2 Lidar Conflation

We followed the same conflation approach as proposed by Stipek et al. [37]. Conflating two datasets collected at differing temporal scales can prove to be problematic due to periods of growth exhibited by the area-of-interest. Therefore, we followed a strict one-to-one relationship requirement when conflating the lidar footprints to both the Microsoft and ORNL footprint datasets (Fig. 2). Please note this conflation process ended with a different number of matched footprints. For example, in Albany, the matching one-to-one footprints for lidar to Microsoft were 108,107 and 116,518 for ORNL (Table 1).

3.3 Morphology Features

We utilized morphology features generated from vector geospatial polygon layers at a building-by-building level [18]. Morphology features have been leveraged to infer building use type, building height, among others [1, 6, 25, 37]. The morphology feature set consists of three types of features: geometric, engineered, and contextual (Table S2). Geometric are basic measures of geometry like area or perimeter. Engineered features describe more complex ideas like compactness or complexity. Contextual features describe the building and its relationship to its neighbors, both spatially and in size. The contextual features are generated at five different scales: 50m, 100m, 250m, 500m, and 1000m. Overall, there are 65 features generated with table S2 providing a description of each feature. The morphology feature set was generated for the ORNL and Microsoft footprints which were selected during the conflation process. We also compared the morphology features generated for each of the 10 cities between the ORNL and Microsoft footprints to better understand the differences between the two footprint datasets.

3.4 Feature Selection

All buildings less than 2m in height were removed from both datasets, a common practice when inferring height from 2D features [25, 37]. Feature reduction was then performed for both datasets via a recursive feature elimination (RFE). This iterative function removes features that display lower significance in relation to the target variable [15]. The features selected for by the RFE are bolded in table S2.

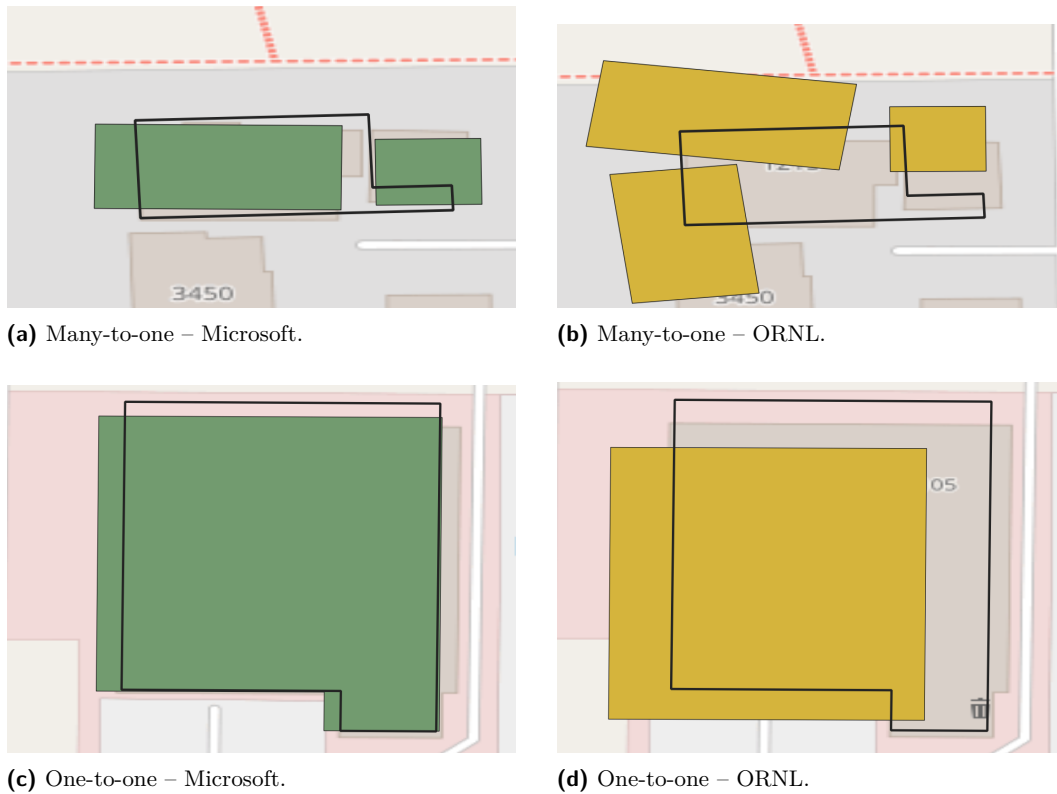
⁷ <https://www.microsoft.com/en-us/maps/bing-maps/building-footprints>

⁸ <https://gisfema.hub.arcgis.com/pages/usa-structures>

■ **Table 1** Built Environment Metrics.

Metric	Microsoft									
	Albany	Boise	Boston	Houston	Nashville	Omaha	Phoenix	Portland	Seattle	Topeka
Count	108,107	64,935	177,997	215,552	50,115	162,180	469,485	104,307	40,169	49,739
Mean	5.50 m	4.16 m	6.94 m	4.28 m	4.63 m	4.96 m	3.85 m	4.86 m	5.05 m	4.39 m
Median	5.33 m	3.83 m	6.89 m	3.74 m	4.27 m	4.77 m	3.48 m	4.47 m	4.53 m	4.09 m
Std	1.54 m	1.24 m	1.95 m	1.72 m	1.52 m	1.30 m	1.24 m	1.71 m	2.39 m	1.26 m
25%	4.41 m	3.34 m	5.61 m	3.40 m	3.74 m	4.14 m	3.15 m	3.68 m	3.77 m	3.59 m
75%	6.33 m	4.75 m	8.04 m	4.65 m	5.09 m	5.55 m	4.17 m	5.64 m	5.71 m	4.89 m
Max	41.16 m	60.77 m	72.69 m	180.90 m	71.22 m	78.32 m	77.07 m	69.71 m	146.45 m	46.32 m
ORNL										
Count	116,518	74,901	198,631	261,225	59,597	181,672	522,666	122,048	51,219	56,038
Mean	5.54 m	4.16 m	7.01 m	4.32 m	4.60 m	4.95 m	3.86 m	4.88 m	5.09 m	4.37 m
Median	5.35 m	3.83 m	6.95 m	3.75 m	4.25 m	4.76 m	3.48 m	4.48 m	4.56 m	4.07 m
Std	1.61 m	1.25 m	2.01 m	1.82 m	1.51 m	1.34 m	1.27 m	1.72 m	2.45 m	1.27 m
25%	4.42 m	3.34 m	5.66 m	3.39 m	3.73 m	4.12 m	3.14 m	3.69 m	3.77 m	3.58 m
75%	6.39 m	4.76 m	8.14 m	4.70 m	5.05 m	5.55 m	4.18 m	5.65 m	5.80 m	4.88 m
Max	71.00 m	60.77 m	101.82 m	180.90 m	71.22 m	78.32 m	96.90 m	69.71 m	146.45 m	46.32 m

Descriptive statistics for the cities selected in this research with Microsoft buildings on the top panel and ORNL footprints on the bottom panel. Please note the differing number of total buildings based on the one-to-one conflation method. For example, the count of buildings in Albany is 108,107 for Microsoft with 116,518 for ORNL.



■ **Figure 2** Examples of footprints which were disqualified due to being a many-to-one (top panel), and footprints which were included in the datasets based on the one-to-one (bottom panel) conflation method. Please note that Microsoft footprints are green, ORNL footprints are yellow, with the ground-truth lidar footprints being the overlaid black outline.

3.5 Model Development

We applied 4 distinct models during this research and compared to a baseline metric, the median, over which any model would be an improvement. The first model we applied, a Linear Regression (LR) model provides a baseline initial estimate and works by assuming there is a linear relationship between the target, height, and the training datum [19]. We next applied a Random Forest (RF) algorithm, first introduced by Breiman [7]. The RF model is a collection of tree-structured classifiers with each tree within a defined forest coming to a decision independent of the other trees. After each tree has inferred a decision based on a random subset of the training datum, a decision is then made for inference based on a majority vote from the individual trees. The XGBoost Regressor (XGB) is a gradient boosting trees algorithm in which decision trees are iteratively added and learn from the previous tree in order to minimize error[11]. This allows for the XGB to learn from each successive tree such that the model will reduce error and improve overall model performance. TabNet, a novel high-performance deep learner designed to help improve tabular datum predictions was also applied [3]. TabNet has been shown to improve run-time and display comparable results to other gradient boosting algorithms [22]. This is the first time, to the authors knowledge, that a deep learning framework has been applied to infer building height from tabular datum. Each model was constructed for each of the cities for both the Microsoft and ORNL footprint datasets and the ensuing steps were taken for each city within both datasets, totalling 20 iterations.

Each city was split into a training and testing dataset, with 70% used for training and 30% for testing. We then leveraged Bayesian optimization using the Hyperopt library [5]. The Bayesian optimization utilizes a prior set of hyper-parameters to inform the successive set of hyper-parameters for testing. It iterates through this process and once complete, it produces the optimum hyper-parameters from a pre-defined grid search space in relation to the lowest RMSE. We selected the following hyper-parameters to fine-tune through the Bayesian optimization: *number of estimators*, *max depth*, *gamma*, *reg alpha*, *reg lambda*, *colsample bytree*, *min child weight*, and *learning rate*. To validate our results, we conducted a 10 fold cross validation (CV) over the entirety of the datum for each individual city. Please note that all references to the XGB RMSE are in reference to the CV score.

3.6 Out of Sample Validation

While conducting a 10-fold CV, we acknowledge that when working with spatially diverse datum, validation should also be applied to distinct geographic areas [24]. To account for this, we conducted a spatial validation similar to that done by Metzger et al. [24] and Stipek et al. [37] in which we randomly selected 3 cities (Albany, Houston, Seattle) as hold-out validation cities for testing for both datasets (ORNL, Microsoft) (Figure S1).

4 Results

The XGB model was chosen based on its superior performance in relation to the other models applied (LR, RF, TabNet). While we acknowledge that the RF outperformed the XGB in certain cities, the XGB was more consistent across both datasets (ORNL, Microsoft). All the results in the following sections are the inferred values from the XGB model. For the results associated with the LR, RF, and TabNet models please see Supplementary Table S3.

4.1 Microsoft Footprints

Each of the 10 cities modeled within the Microsoft dataset showed improvement upon the median value generated for both the MAE and RMSE (Table 2). The median values generated present a baseline value for building height over which any improvement in relation to MAE or RMSE can be considered an improvement over a baseline estimate. Phoenix showed the highest goodness of fit, R^2 , with a metric of 61% with Nashville displaying the lowest, 39%. In relation to improvement upon the median RMSE baseline, Seattle was the highest, with $-0.59m$ improvement, going from the median of $2.44m$ to a modeled output RMSE of $1.85m$. Topeka and Albany displayed the lowest improvement, displaying differences of $-0.29m$ and $-0.35m$, respectively. On average, the percentage improvement across the 10 cities from the median RMSE to the modeled RMSE was 32%.

4.2 ORNL Footprints

The modeled ORNL footprints also showed improvement upon the median MAE and RMSE for each of the 10 cities (Table 2). Phoenix displayed the highest R^2 score, 60%, with Seattle displaying the lowest, with a score of 38%. For improvement upon the RMSE baseline, Seattle showed the highest improvement, $-0.58m$, with Topeka displaying the lowest, $-0.30m$. The average percent improvement from the median RMSE to the modeled RMSE across the 10 cities was 31%.

■

Table 2 Microsoft and ORNL Results.

Metric	Microsoft									
	Albany	Boise	Boston	Houston	Nashville	Omaha	Phoenix	Portland	Seattle	Topeka
Median MAE	1.12 m	0.84 m	1.45 m	0.92 m	0.91 m	0.91 m	0.78 m	1.15 m	1.30 m	0.82 m
Median RMSE	1.55 m	1.28 m	1.95 m	1.80 m	1.56 m	1.32 m	1.30 m	1.75 m	2.44 m	1.29 m
XGBoost MAE	0.82 m	0.58 m	1.00 m	0.64 m	0.69 m	0.60 m	0.45 m	0.82 m	1.01 m	0.58 m
XGBoost RMSE	1.20 m	0.89 m	1.47 m	1.21 m	1.17 m	0.95 m	0.80 m	1.27 m	1.85 m	1.00 m
XGBoost R^2	40%	46%	45%	48%	39%	51%	61%	48%	42%	41%
% Improvement	25%	36%	28%	39%	29%	33%	48%	32%	28%	25%
ORNL										
Median MAE	1.15 m	0.85 m	1.48 m	0.96 m	0.90 m	0.91 m	0.79 m	1.16 m	1.33 m	0.82 m
Median RMSE	1.63 m	1.30 m	2.02 m	1.90 m	1.55 m	1.35 m	1.33 m	1.77 m	2.49 m	1.31 m
XGBoost MAE	0.85 m	0.61 m	1.03 m	0.68 m	0.72 m	0.63 m	0.49 m	0.86 m	1.02 m	0.60 m
XGBoost RMSE	1.26 m	0.91 m	1.49 m	1.33 m	1.24 m	0.98 m	0.81 m	1.30 m	1.91 m	1.01 m
XGBoost R^2	39%	44%	43%	46%	41%	47%	60%	42%	38%	40%
% Improvement	26%	35%	30%	35%	22%	32%	49%	31%	26%	26%

Please note that the top panel are model results from the Microsoft footprints with the model results derived from ORNL footprints on the bottom panel. For the metrics, we display the MAE and RMSE associated with the median before displaying the model (XGBoost) results below. The median results are the baseline over which any model is an improvement over the simplest possible method in relation to inferring building height. We also display the % difference between the CV RMSE and median RMSE to showcase the improvement upon the baseline.

Across all 10 cities, the largest difference between the Microsoft and ORNL footprints in relation to RMSE was $0.09m$, observed in both Albany and Nashville with the lowest difference observed being $0.01m$ in Topeka. In relation to the percentage improvement upon the median baseline for RMSE when comparing Microsoft and ORNL, the largest difference in improvement was 7% (29% - Microsoft, 22% - ORNL), observed in Nashville with six of the cities showing only 1% difference.

4.3 Morphology Differences

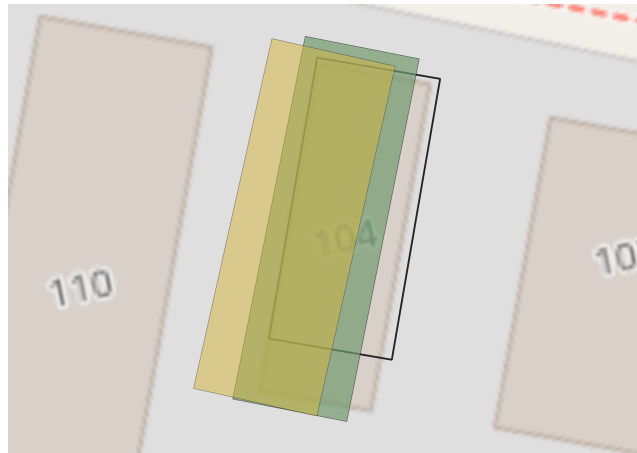
When comparing the differences between the morphology features generated for the ORNL and Microsoft footprints, the majority of the features showed minimal differences. However, there were some features which displayed differences, primarily the contextual and engineered features such as *complexity ps*, *n count*, and *n size mean* (Table S4). For the Microsoft footprints, the *complexity ps* displayed a median of 2,812 while the ORNL median was 7,717, signifying differences within the shapes of the footprints (Table S4). For the *n count 500*, Microsoft displayed a max count of 1,267, compared to 928 for ORNL, signifying a difference in the number of footprints within a 500m radius. For *n size mean 500*, the max feature displayed by Microsoft was 210,048 with a value of 102,230 by ORNL, highlighting the differences in footprint sizes within a 500m radius. These results are similar to the research conducted by Chamberlain et al. ([8]) and Gonzales ([14]).

4.4 Out of Sample Validation

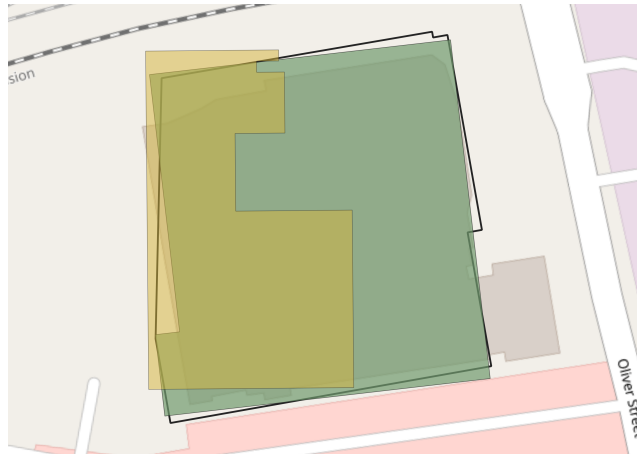
When testing on Microsoft footprints, Albany did not improve upon the median RMSE as the XGB RMSE displayed a value of $1.55m$ and a R^2 score of -1% (Table S5). However, the other cities which were tested with Microsoft footprints showed improvements upon the median RMSE, being $-0.20m$ in Houston and $-0.35m$ in Seattle. All three of the cities when tested on ORNL footprints displayed improvements upon the median RMSE, being $-0.07m$ for Albany, $-0.22m$ for Houston and $-0.37m$ for Seattle.

5 Discussion

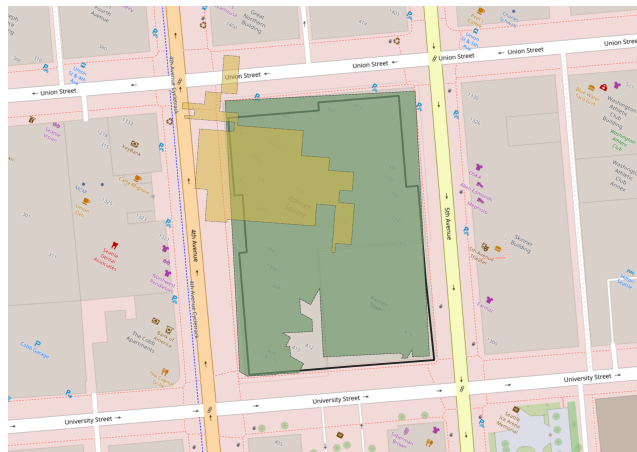
The expansion of open-source building footprint datasets has provided the possibility for leveraging these products to characterize the built environment. Our results show that, across 3.09 million buildings in the U.S., our method of inferring height from footprint information alone is effective for datasets produced by ORNL as well as the much larger and globally available dataset from Microsoft. Furthermore, our height prediction process is reliable and agnostic to building footprint source. This finding ensures that our approach of inferring height from footprint morphology features can be scaled to leverage other publicly available footprints, such as Microsoft footprints. By inferring building height, this method provides valuable contextual information for population density estimation, building energy, disaster management, and UHI's [10, 12, 31, 32, 39].



(a) Boston.



(b) Houston.



(c) Seattle.

■ **Figure 3** Here we display the different shapes in footprints displayed by Microsoft (green), ORNL (yellow) with the lidar footprints a black outline. For Boston (a), the lidar height is $6.90m$, with our prediction based on the MS footprint being $8.52m$ and $7.82m$ inferred on the ORNL footprint. In Houston (b), the lidar height is $6.19m$, our prediction based on the MS footprint is $10.29m$ and $9.28m$ with the ORNL footprint. In Seattle (c), the lidar height is $38.02m$, the height inferred on the MS footprint is $18.11m$ and $50.86m$ on the ORNL footprint.

While the main objective of this research is to test the efficacy of leveraging open-source footprints, we applied various models to ensure the best possible method was selected. It is important to note that the TabNet model did not outperform either the RF or XGB for any cities across both footprint sources. In some instances, such as in Phoenix, the difference was -18% in relation to R^2 results displayed by the XGB. However, in other cities (Houston), the TabNet outperformed the percent improvement displayed by the RF by +1% in relation to R^2 . Regardless, the tree based approach consistently outperformed the TabNet model which signifies that while deep learning models developed for tabular data have made progress [21, 22], in this instance tree based models show higher accuracy.

While successful, our study does have limitations that need to be acknowledged. First, the area-of-interest is only within one country, the U.S., and more work is needed to expand this approach to additional countries. It is known that the built environment varies both spatially and temporally and a more diverse sample set is needed to further validate this approach [6, 25, 27, 34, 37]. Another limitation is that during our strict one-to-one conflation process, building footprints that don't have a one-to-one match are removed and therefore not included in the morphology feature generation. Contextual features that look at a building's neighbors have been found to be influential to the model's behaviour and therefore, the model may not perform as well on the filtered dataset as it would on the unfiltered dataset [37]. For example, the generated feature *n count* measures the number of centroids within a defined radius surrounding a building. This was evidenced by the range of values displayed for the *n count 500* for the ORNL when compared to Microsoft (Table S4). The range in values for *n count 500* signifies that at a 500 m radius, there are differences associated with the total number of buildings, which can influence the model's ability to infer an individual building's height.

Additionally, there needs to be a formal analysis completed to understand if it is possible to train on one distinct footprint dataset and test on another. For example, due to the differences discussed between the Google and Microsoft datasets within Africa [8, 14], can it be possible to train on the Google footprints to then infer height on the Microsoft footprints. While the approach presented in this research shows the ability to infer height agnostic of footprint source, it does not test across the sources, i.e. training on ORNL and testing on Microsoft.

Furthermore, the differences in footprint shape based on the pixel identification and regularization process can lead to irregularities in predicted height (Fig. 3). For example, the complexity ratio, an engineered feature that is the shape length divided by the shape area which shows high significance in relation to inferring height, can vary depending on the footprint shape [37]. For example, in Boston, we display the Microsoft, ORNL and lidar footprint for one specific building where the footprint shapes are similar and the height prediction for the MS footprint is 8.52m and 7.82m for the ORNL footprint (Fig. 3). However, when there are differences displayed by the building's footprint, there can be large differences associated with the predicted height, as evidenced in the example in Seattle where the height inferred from the Microsoft footprint is 18.11m and 50.86m on the ORNL footprint. Therefore, based on the shape and size of the footprint, the inferred height may vary, as displayed in figure 3. While the majority of the morphology features showed minimal differences in their distributions, it is important to note that some of the engineered features, such as *complexity ps* showed differences (Table S4). Therefore, while the approach presented in this research has proven it is possible to infer height from various footprint sources, it would be irresponsible to apply without additional testing if leveraging an additional footprint source, such as Google.

This research has highlighted the need for multiple avenues of future work. A comprehensive analysis in relation to the distributions displayed by the morphological features is necessary to truly understand the differences displayed between ORNL and Microsoft datasets. As the scope of this paper is to investigate if it is possible to infer height from both products, we do not fully investigate the differences displayed by the ORNL and Microsoft footprints in relation to the engineered and contextual features. Other potential work could explore the possibility of training on one homogeneous footprint data source and testing on another.

6 Conclusion

In this paper, we demonstrate the ability of our method to infer height from building footprints derived from different sources (ORNL, Microsoft). Our results show that, across over 3 million footprints in the U.S., we successfully infer building height within 3m of the ground truth height with 98% accuracy. More importantly, while previous work has proven that it is possible to infer height from footprints derived from satellite datum, this is the first time, to our knowledge, that a comparison study has been completed that indicates a machine-learning height inference method can be applied across multiple datasets. We believe our approach is successful due to the ability to learn from the distinct morphology features, regardless of the footprint dataset. This is a significant finding which displays the generalization of our method to inferring height regardless of how the building footprints are extracted and regularized. Furthermore, this opens the door to now leverage the over 1 billion Microsoft footprints to infer building height at a building-by-building level across the globe.

References

- 1 D. Adams, T. Hauser, and J. Moehl. Decoding ethiopian abodes: Towards classifying buildings by occupancy type using footprint morphology. *22nd IEEE International Conference on Machine Learning and Applications*, 2023.
- 2 W. Alsabahn, T. Alotaiby, and B. Dudin. Detecting buildings and nonbuildings from satellite images using u-net. *Computational Intelligence and Neuroscience*, 2022. doi:10.1155/2022/4831223.
- 3 S. Arik and T. Pfister. Tabnet: Attentive interpretable tabular learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35, 2021. doi:10.1609/aaai.v35i8.16826.
- 4 C. Ayala, R. Sesma, C. Aranda, and M. Galar. A deep learning approach to an enhanced building footprint and road detection in high-resolution satellite imagery. *Remote Sensing*, August 2021. doi:10.3390/rs13163135.
- 5 J. Bergstra, B. Komer, C. Eliasmith, D. Yamins, and D. Cox. Hyperopt: a python library for model selection and hyperparameter optimization. *Computational Science & Discovery*, 8:1, 2015. doi:10.1088/1749-4699/8/1/014008.
- 6 F. Biljecki, H. Ledoux, and Stoter J. Generating 3d city models without elevation data. *Computers, Environment and Urban Systems*, 64, July 2017. doi:10.1016/j.compenvurbsys.2017.01.001.
- 7 L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. doi:10.1023/A:1010933404324.
- 8 H. Chamberlain, E. Darin, A. Adewole, W. Jochem, A. Lazar, and A. Tatum. Building footprint data for countries in africa: to what extent are existing data products comparable? *Computers, Environment and Urban Systems*, 110, 2024. doi:10.21203/rs.3.rs-3334423/v1.

- 9 C. Chawda, J. Aghav, and S. Udar. Extracting building footprints from satellite images using convolutional neural networks. *2018 International Conference on Advances in Computing, Communications and Informatics*, 2018. doi:10.1109/ICACCI.2018.8554893.
- 10 F. Chen, H. Kusaka, R. Bornstein, J. Ching, C.S.B. Grimmond, S. Grossman-Clarke, T. Loidan, K. Manning, A. Martilli, S. Miao, D. Sailor, F. Salamanca, H. Taha, M. Tewari, X. Wang, A. Wyszogrodzki, and C. Zhang. The integrated wrf/urban modelling system: development, evaluation, and applications to urban environmental problems. *International Journal of Climatology*, 31, 2011. doi:10.1992/joc.2158.
- 11 T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016. doi:10.1145/2939672.2939785.
- 12 Y. Chen, T. Hong, X. Luo, and B. Hooper. Development of city buildings dataset for urban building energy modeling. *Energy and Buildings*, 183, 2019. doi:10.1016/j.enbuild.2018.11.008.
- 13 A. Comber, M. Umezaki, R. Zhou, Y. Ding, Y. Li, F. Hua, H. Jiang, and A. Tewkesbury. Using shadows in high-resolution imagery to determine building height. *Remote Sensing Letters*, 3, December 2011. doi:10.1080/01431161.2011.635161.
- 14 J. Gonzales. Building-level comparison of microsoft and google open building footprints datasets. *12th International Conference on Geographic Information Science*, 277, 2023.
- 15 B. Gregorutti, B. Michel, and P. Saint-Pierre. Correlation and variable importance in random forests. *Statistics and Computing*, 27, March 2017. doi:10.1007/s11222-016-9646-1.
- 16 A. Gribov. Optimal compression of a polyline while aligning to preferred directions. *2019 International Conference on Document Analysis and Recognition*, 2019.
- 17 United Nations Habitat. Urbanization and development: Emerging futures. *Nairobi, Kenya: UN Habitat*, 2016.
- 18 T. Hauser, J. Moehl, E. Schmidt, B. Morris, D. Adams, and H. L. Yang. Usa structures phase 2 technical report. Technical report, Oak Ridge National Laboratory, August 2023. doi:10.2172/2076189.
- 19 G. James, D. Witten, T. Hastie, and R. Tibshirani. Linear regression. *An Introduction to Statistical Learning. Springer Texts in Statistics.*, 12, 2021. doi:10.1007/978-1-0716-1418-1_3.
- 20 W. Jochem, D. Leasure, O. Pannell, H. Chamberlain, P. Jones, and A. Tatem. Classifying settlement types from multi-scale spatial patterns of building footprints. *Environment and Planning B: Urban Analytics and City Science*, 48, 2020. doi:10.1177/2399808320921208.
- 21 R. Kanasz, P. Drotar, P. Gnyp, and M. Zoricak. Clash of titans on imbalanced data: Tabnet vs xgboost. *2024 IEEE Conference on Artificial Intelligence*, 2024. doi:10.1109/CAI59869.2024.00068.
- 22 A. Lewandowska. Xgboost meets tabnet in predicting the costs of forwarding contracts. *17th Conference on Computer Science and Intelligence Systems (FedCSIS)*, 2022. doi:10.15439/2022F294.
- 23 W. Li, C. He, J. Fang, H. Zheng, J. nd Fu, and L. Yu. Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source gis data. *remote sensing*, 11, February 2019. doi:10.3390/rs11040403.
- 24 N. Metzger, J. Vargas-Munoz, R. Daudt, K. Kellenberger, T. Ton-That Whelan, M. Imran, K. Schindler, and D. Tuia. Fine-grained population mapping from coarse census counts and open geodata. *Scientific Reports*, 12, 2022. doi:10.1038/s41598-022-24495-w.
- 25 N. Milojevic-Dupont, N. Hans, L. Kaack, M. Zumwald, F. Andrieux, D. Soares, S. Lohrey, P. Pichler, and F. Creutzig. Learning from urban form to predict building heights. *PLoS One*, 15:12, December 2020. doi:10.1371/journal.pone.0242010.
- 26 A. Milosavljevic. Automated processing of remote sensing imagery using deep semantic segmentation: A building footprint extraction case. *International Journal of Geo-Information*, August 2020. doi:10.3390/ijgi9080486.

- 27 F. Nachtigall, N. Milojevic-Dupont, F. Wagner, and F. Creutzig. Predicting building age from urban form at large scale. *International Journal of Environmental Research and Public Health*, 105, October 2023. doi:10.1016/j.compenvurbsys.2023.102010.
- 28 V. Oludare, L. Kezebou, K. Panetta, and S. Agaian. Semi-supervised learning for improved post-disaster damage assessment from satellite imagery. *Proceedings from Multitmodal Image Exploitation and Learning*, 11734, 2021. doi:10.1117/12.2586232.
- 29 Z. Pan, J. Xu, Y. Guo, Y. Hu, and G. Wang. Deep learning segmentation and classification for urban village using a worldview satellite image based on u-net. *Remote Sensing*, 12, 2020. doi:10.3390/rs12101574.
- 30 K. Reda and M. Kedzierski. Detection, classification and boundary regularization of buildings in satellite imagery using faster edge region convolutional neural networks. *Remote Sensing*, July 2020. doi:10.3390/rs12142240.
- 31 C. Robinson, F. Hohman, and B. Dilkina. The integrated wrf/urban modelling system: development, evaluation, and applications to urban environmental problems. *1st ACM SIGSPATIAL Workshop on Geospatial Humanities*, 2017. doi:10.1145/3149858.3149863.
- 32 A. Rodriguez and J. Wegner. Counting the uncountable: Deep semantic density estimation from space. *Pattern Recognition*, 11269, 2019. doi:10.1007/978-3-030-12939-2_24.
- 33 O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention*, 2015.
- 34 Q. Shi, J. Zhu, Z. Liu, H. Guo, S. Gao, M. Liu, Z. Liu, and X. Liu. The last puzzle of global building footprints—mapping 280 million buildings in east asia based on vhr images. *Journal of Remote Sensing*, 4, May 2024. doi:10.34133/remotesensing.0138.
- 35 W. Sirko, S. Kashubin, M. Ritter, A. Annkah, Y. Bouchareb, Y. Dauphin, D. Keysers, M. Neumann, M. Cisse, and J. Quinn. Continental-scale building detection from high resolution satellite imagery. *arXiv*, July 2021. doi:10.48550/arXiv.2107.12283.
- 36 H. Song, L. Yang, and J. Jung. Self-filtered learning for semantic segmentation of buildings in remote sensing imagery with noisy labels. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16, 2023. doi:10.1109/JSTARS.2022.3230625.
- 37 C. Stipek, T. Hauser, D. Adams, J. Epting, C. Brelford, J. Moehl, P. Dias, J. Piburn, and R. Stewart. Inferring building height from footprint morphology data. *Nature: Scientific Reports*, 14, August 2024. doi:10.1038/s41598-024-66467-2.
- 38 P. Ulmas and I. Liiv. Segmentation of satellite imagery using u-net models for land cover classification. *arXiv*, 2020. doi:10.48550/arXiv.2003.02899.
- 39 N. Wang, S. Goel, and A. Makhmalbaf. Commercial building energy asset score program overview and technical protocol. *Technical Report, PNNL-22045 Rev 1.1*, 2013.
- 40 S. Wei, S. Ji, and M. Lu. Toward automatic building footprint delineation from aerial images using cnn and regularization. *IEEE Transactions on Geoscience and Remote Sensing*, March 2020. doi:10.1109/TGRS.2019.2954461.
- 41 L. Yang, M. Laverdiere, T. Hauser, B. Swan, E. Schmidt, J. Moehl, A. Reith, D. Adams, B. Morris, J. McKee, M. Whitehead, and M. Tuttle. A baseline inventory with critical attribution for the us and its territories. *Nature: Scientific Data*, 11, 2024. doi:10.1038/s41597-024-03219-x.
- 42 L. Yang, D. Yuan, J. nd Lunga, M. Laverdiere, A. Rose, and B. Bhaduir. Building extraction at scale using convolutional neural network: Mapping of the united states. *IEEE Journal of Selected Topics in Applied Earth Observation and Remote Sensing*, 11, July 2018. doi:10.1109/JSTARS.2018.2835377.
- 43 K. Zhao, M. Kamran, and G. Sohn. Boundary regularized building footprint extraction from satellite images using deep neural networks. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2020. doi:10.5194/isprs-annals-V-2-2020-617-2020.

A **Supplementary Tables**

Table S1 Temporality of lidar, ORNL footprints, and Microsoft footprints.

Location	Lidar	ORNL	Microsoft
Albany, NY, USA	10/9/2012	19/10/2019	26/4/2023
Boise, ID, USA	8/3/2013	4/8/2018	26/4/2023
Boston, MA, USA	20/5/2009	9/11/2019	26/4/2023
Houston, TX, USA	22/1/2010	21/10/2021	26/4/2023
Nashville, TN, USA	6/6/2006	6/6/2019	26/4/2023
Omaha, NE, USA	24/4/2013	20/5/2020	26/4/2023
Phoenix, AZ, USA	4/10/2014	27/2/2020	26/4/2023
Portland, OR, USA	20/9/2010	27/2/2020	26/4/2023
Seattle, WA, USA	6/5/2010	3/3/2020	26/4/2023
Topeka, KS, USA	10/12/2008	23/11/2020	26/4/2023

Please note that the dates for the footprint sources are in DD/MM/YYYY format.

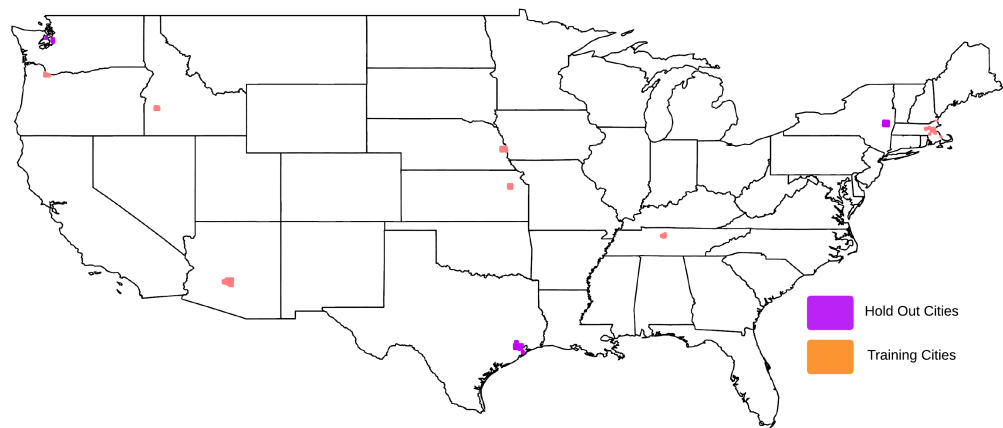


Figure S1 Map which displays the 3 randomly selected hold out cities for our additional validation step.

■ **Table S2** Building Morphology Features.

Feature	Description
<i>Geometric Features</i>	
shape area	Area of polygon in un-projected units
shape length	Perimeter length in un-projected units
sqft	Area in square feet
sqmeters	Area in square meters
lat dif	Maximum latitude minus minimum latitude in un-projected units
long dif	Maximum longitude minus minimum longitude in un-projected units
envel area	Area of bounding box of geometry in un-projected units
vertex count	Count of vertices in geometry
geom count	Count of polygons in the geometry
<i>Engineered Features</i>	
complexity ratio	Shape length / shape area
iasl	Inverse average segment length
vpa	Vertices per area
complexity ps	Complexity per segment, average complexity within each segment
ipq	Isoperimetric quotient, shape area maximization for given perimeter length
<i>Contextual Features</i>	
n count*	Number of building centroids within a given distance
omd*	Observed mean distance from building within a given distance
emd*	Expected mean distance from building within a given distance
nnd*	Nearest neighbor distance from building
nni*	Nearest neighbor index, overall pattern of points within a given distance
intensity*	Amount of nni occurring
n size mean*	Average size of buildings within a given distance
n size std*	Standard deviation of buildings within a given distance
n size min*	Smallest building size within a given distance
n size max*	Largest building size within a given distance
n size cv*	Coefficient of variation of building size within a given distance

* Denotes feature being calculated on multiple scales. Bolded features highlight the features selected for use.

Table S3 Microsoft and ORNL Results.

	Microsoft									
Metric	Albany	Boise	Boston	Houston	Nashville	Omaha	Phoenix	Portland	Seattle	Topeka
Median MAE	1.12 m	0.84 m	1.45 m	0.92 m	0.91 m	0.91 m	0.78 m	1.15 m	1.30 m	0.82 m
Median RMSE	1.55 m	1.28 m	1.95 m	1.80 m	1.56 m	1.32 m	1.30 m	1.75 m	2.44 m	1.29 m
LR MAE	1.08 m	0.75 m	1.21 m	0.91 m	0.85 m	0.83 m	0.69 m	1.09 m	1.19 m	0.77 m
LR RMSE	1.44 m	1.08 m	1.67 m	1.57 m	1.38 m	1.14 m	1.04 m	1.56 m	2.16 m	1.14 m
LR R^2	12%	25%	27%	22%	21%	19%	28%	19%	26%	17%
% Improvement	7%	16%	14%	13%	12%	14%	20%	11%	11%	12%
RF MAE	0.84 m	0.60 m	1.00 m	0.67 m	0.71 m	0.61 m	0.47 m	0.85 m	1.01 m	0.58 m
RF RMSE	1.20 m	0.94 m	1.46 m	1.32 m	1.23 m	0.91 m	0.79 m	1.27 m	1.94 m	0.96 m
RF R^2	39%	44%	44%	44%	37%	48%	58%	46%	40%	41%
% Improvement	23%	27%	25%	27%	21%	31%	39%	27%	20%	26%
TabNet MAE	0.94 m	0.65 m	1.12 m	0.77 m	0.80 m	0.68 m	0.56 m	0.95 m	1.14 m	0.65 m
TabNet RMSE	1.30 m	0.93 m	1.58 m	1.29 m	1.34 m	1.02 m	0.92 m	1.38 m	2.28 m	1.03 m
TabNet R^2	27%	40%	35%	40%	28%	38%	42%	34%	27%	31%
% Improvement	16%	27%	19%	28%	14%	23%	29%	21%	7%	20%
ORNL										
Median MAE	1.15 m	0.85 m	1.48 m	0.96 m	0.90 m	0.91 m	0.79 m	1.16 m	1.33 m	0.82 m
Median RMSE	1.63 m	1.30 m	2.02 m	1.90 m	1.55 m	1.35 m	1.33 m	1.77 m	2.49 m	1.31 m
LR MAE	1.10 m	0.75 m	1.27 m	0.90 m	0.83 m	0.85 m	0.70 m	1.12 m	1.21 m	0.78 m
LR RMSE	1.51 m	1.07 m	1.75 m	1.37 m	1.30 m	1.20 m	1.07 m	2.04 m	2.34 m	1.14 m
LR R^2	13%	26%	23%	34%	24%	18%	29%	-32%	23%	20%
% Improvement	7%	18%	13%	28%	16%	11%	20%	-15%	6%	13%
RF MAE	0.84 m	0.60 m	1.01 m	0.66 m	0.70 m	0.60 m	0.47 m	0.85 m	1.02 m	0.56 m
RF RMSE	1.26 m	0.91 m	1.47 m	1.19 m	1.18 m	0.93 m	0.80 m	1.34 m	2.18 m	0.93 m
RF R^2	39%	46%	46%	50%	38%	51%	60%	43%	33%	48%
% Improvement	23%	30%	27%	37%	24%	31%	40%	24%	12%	29%
TabNet MAE	0.94 m	0.67 m	1.15 m	0.76 m	0.79 m	0.71 m	0.56 m	0.94 m	1.10 m	0.65 m
TabNet RMSE	1.35 m	0.99 m	1.58 m	1.24 m	1.24 m	1.04 m	0.90 m	1.47 m	2.08 m	1.00 m
TabNet R^2	31%	36%	37%	46%	30%	38%	49%	31%	39%	38%
% Improvement	17%	24%	22%	35%	20%	23%	32%	17%	16%	24%

Results for the LR, RF and TabNet models for the Microsoft (top) and ORNL (bottom) footprints.

■ **Table S4** Morphology Differences – Microsoft and ORNL Footprints.

	Microsoft Footprints		
Metrics	<i>Complexity PS</i>	<i>N Count 500</i>	<i>N Size Mean 500</i>
Mean	2,767	354	2,435
Median	2,812	295	2,029
Std	1,005	236	2,030
Min	47	2	775
25%	2,025	180	1,742
75%	3,496	487	2,457
Max	15,617	1,267	210,048
	ORNL Footprints		
Mean	7,673	289	2,124
Median	7,717	252	1,697
Std	2,405	176	1,779
Min	56	1	848
25%	6,240	158	1,402
75%	9,099	398	2,175
Max	21,160	928	102,320

The morphology features displayed in this table were generated in Albany, one of the 10 cities investigated during this research.

■ **Table S5** Out of Sample Validation Results.

	Microsoft Footprints		
Cities	<i>Albany</i>	<i>Houston</i>	<i>Seattle</i>
Median MAE	1.12 m	0.92 m	1.30 m
Median RMSE	1.55 m	1.80 m	2.44 m
XGB MAE	1.12 m	0.97 m	1.20 m
XGB RMSE	1.55 m	1.60 m	2.09 m
XGB R^2	-1%	13%	23%
	ORNL Footprints		
Median MAE	1.15 m	0.96 m	1.33 m
Median RMSE	1.63 m	1.90 m	2.49 m
XGB MAE	1.11 m	1.07 m	1.23 m
XGB RMSE	1.56 m	1.68 m	2.12 m
XGB R^2	6%	14%	24%

The results displayed for our out of sample validation test.

CityJSON Management Using Multi-Model Graph Database to Support 3D Urban Data Management

Muhammad Syafiq ✉ 

3D GIS Research Lab, Universiti Teknologi Malaysia, Johor Bahru, Malaysia

Suhaibah Azri¹ ✉ 

3D GIS Research Lab, Universiti Teknologi Malaysia, Johor Bahru, Malaysia

Uznir Ujang ✉ 

3D GIS Research Lab, Universiti Teknologi Malaysia, Johor Bahru, Malaysia

Abstract

The prevalence of 3D city models in urban applications is increasing due to their lightweight and flexibility, making them adaptable to various applications. However, effective data interoperability remains an issue. Managing 3D city models within a database can improve urban data management applications such as data enrichment and efficient querying. Motivated by the need for better interoperability of 3D city models, this paper proposes a novel method for storing CityJSON using the concept of a multi-model graph database as a foundation for enriching its semantics. The proposed approach involves decomposing CityJSON objects into smaller JSON components, which are then abstracted into graph elements. Parent-child and other relationship attributes are modelled to capture the hierarchical and associative structures of the CityJSON data. A specific programme is employed to preprocess CityJSON data based on several conditions before being loaded into the graph database. Our multi-model approach allows three types of queries: document, graph, and hybrid. The latter combines both document and graph query. Comparative evaluation against relational databases demonstrates that the proposed method outperforms in terms of query performance. The improved query performance is attributed to the advantage of graph database that reduces the need for joins and the ability to efficiently index and navigate JSON data. The findings of this study establish a foundation for semantic enrichment of 3D city models to improve interoperability and support advanced urban data management.

2012 ACM Subject Classification Information systems → Graph-based database models; Information systems → Geographic information systems; Information systems → Database design and models

Keywords and phrases CityJSON, Graph Database, 3D City Model, 3D GIS, Interoperability

Digital Object Identifier 10.4230/LIPICs.GIScience.2025.2

Funding This work is supported by the Ministry of Higher Education through the Fundamental Research Grant Scheme (FRGS/1/2022/WAB07/UTM/02/3). The highest appreciation is offered to UTMNexus scholarship for sponsoring the study of the first author.

Suhaibah Azri: Fundamental Research Grant Scheme (FRGS/1/2022/WAB07/UTM/02/3).

1 Introduction

Urban management reliance on 3D city models has grown steadily in recent years, driven by their role in various urban applications such as energy demand modelling, indoor navigation, and sustainability studies [27]. A virtual representation of city and urban data is required for meeting the demands of modern urban applications to enable effective decision-making, efficient resource allocation, and adept strategic planning [28, 10, 19]. 3D city models are

¹ Corresponding author



© Muhammad Syafiq, Suhaibah Azri, and Uznir Ujang;
licensed under Creative Commons License CC-BY 4.0

13th International Conference on Geographic Information Science (GIScience 2025).

Editors: Katarzyna Sila-Nowicka, Antoni Moore, David O'Sullivan, Benjamin Adams, and Mark Gahegan;
Article No. 2; pp. 2:1–2:15



Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

primarily developed to visualise and represent urban objects [26, 14]. However, they can be stored in a database for advanced querying, analysis, and integration of urban objects with associated urban data.

CityJSON is a 3D city model format encoded in JSON, which serves as an exchange format for CityGML [12]. It carries most of the schema exists in CityGML, allowing it to represent urban objects, such as buildings, bridges, vegetation, and city furniture, in 3D and multiple Level of Details (LoD). CityJSON is also capable of storing attributes based on their data structure. Their data structure adopts a hierarchical and nested design that enables clear parent-child relationships among urban objects. This organisation allows the attributes, geometries, and semantic information to be assigned directly to their corresponding urban objects. However, their structure can get heavily nested where querying information from the whole document would be complex and inefficient. Therefore, decomposing their structure into a more readable and less nested approach is a more intuitive method.

Research into storing 3D city models has explored both relational (RDBMS) and non-relational (NoSQL) databases. In RDBMS, 3D city model components are typically stored in tabular formats, with relationships like parent-child are handled through additional tables. This structure introduces limitations due to the reliance on numerous joins, which reduce efficiency and increase query complexity [16]. RDBMS also lacks native support for hierarchies and inheritance, making it less suited for representing real-world urban objects and their complex relationships [33, 5]. Consequently, relational databases struggle with object-oriented representations and nested structures, leading to inefficiencies in querying 3D city model data [4]. The lack of flexibility to represent 3D urban data in an object-oriented manner is thus open for further research. Moreover, object-oriented approaches have been recognised for effectively modelling complex relationships in 3D GIS, supporting urban applications and enabling detailed structural analyses [33, 21, 11].

NoSQL databases have also been explored as a replacement to address RDBMS limitations, particularly regarding inflexible schema. They are capable of structuring information using object-oriented approach, which allows for modelling hierarchies and inheritance relationships. This capability is highly relevant for managing 3D urban data that involves deep hierarchies and complex information associations [24]. Document-based and graph-based databases have been used to store 3D city models, with data decomposition being a common method of data insertion. CityJSON, which is encoded in the JSON format, further facilitates data insertion into NoSQL databases like MongoDB and ArangoDB as they readily accept data in JSON format. This compatibility reduces the need for data format conversion as CityJSON can be stored directly in its native format. However, storing CityJSON as a whole before unnesting its components is cumbersome as it will further complicate querying and analysis.

This study addresses the limitations of storing CityJSON in relational databases, particularly the challenges of handling its nested structure and object-oriented modelling, by using ArangoDB, a multi-model graph database. CityJSON components are decomposed and stored as documents, which also serve as nodes in a Labelled Property Graph (LPG) structure. Relationships are modelled to reflect parent-child hierarchies and the inheritance of geometry and semantic attributes, while attributes themselves are represented as edges linked to their respective City Objects. This graph-based transformation emphasises semantic decomposition over geometric detail and offers improved query performance compared to relational models.

2 Related Works

CityJSON is a lightweight 3D city model exchange format for CityGML designed to enhance the interoperability of 3D city models. Its JSON-based encoding simplifies storage and parsing compared to the XML-based CityGML format, which is more verbose and often complicated to handle. The lightweight nature of CityJSON makes it preferred by programmers due to reduced complexity when building applications around it [12].

Furthermore, JSON is a human and machine-readable format that simplifies the process of data manipulation and information retrieval. This makes CityJSON more accessible for integration with web applications, APIs, and databases that inherently support JSON. CityJSON widespread compatibility reduces some challenges in its various applicability, which ultimately improves the usability and interoperability of 3D city models across many urban applications. This practical advantage underlines CityJSON as a logical, more intuitive choice to improve the interoperability and utility of 3D city model data.

When storing information in a DBMS, relational databases are typically the first choice following their widespread use and more functionality. 3DCityDB [29] is a relational database schema designed specifically to store OGC-standard 3D city models. It is built based on spatially enhanced relational databases of either PostgreSQL with PostGIS extension or Oracle Spatial. 3DCityDB provides several functionalities including storing, managing, visualising, analysing, and exporting 3D city models in CityGML format. The initial development of 3DCityDB did not offer support for CityJSON; however, subsequent developments introduced the capability to import and export CityJSON. Another relational database solution to store CityJSON is CityREST [13], which is a RESTful API designed to stream CityJSON datasets over the web. It is built on top of PostgreSQL, which offers several key functionalities like retrieving city objects, filtering city objects within a specified bounding box, and data filtering. A more recent approach is CJDB, which is a relational database schema designed for storing CityJSON built on top of PostgreSQL. It is developed as a more efficient alternative to the 3DCityDB schema for storing and managing CityJSON data. Unlike 3DCityDB, CJDB significantly reduces the large number of tables required to store similar datasets. This design simplifies data management, which in turn reduces memory usage [17].

Concerns have been raised by [1] towards the unsuitability of RDBMS for storing OGC standard data models due to the risk of impedance mismatch. This issue arises when attempting to map object-oriented data models into relational schemas can potentially lead to the loss of critical information or relationships. Despite its extended capabilities, ORDBMS still depends on table joins, making it less suited for modelling the hierarchical structure of 3D city models. For instance, representing parent-child relationships requires additional tables and duplicating City Objects as foreign keys, which increases complexity and reduces performance. In contrast, LPG-based graph databases handle such hierarchies more intuitively by directly linking nodes without duplication, offering better node reusability and efficiency. Moreover, storing and querying JSON data in relational databases involves specialised operations that degrade performance as data size and complexity grow. Furthermore, the storage and querying of JSON-based data in relational databases require specialised operations. These operations may impact query performance, particularly as the complexity and size of the data increase.

Existing solutions point towards the use of NoSQL databases that are object-oriented to store information and relationships more efficiently than relational database that lacks the schema and flexibility to represent a real-world entity [8]. The process of locating object-oriented data in the tabular format relevant to relational databases can be difficult and

prone to misrepresentation of information [15]. This limitation arises because of relational databases that are not designed to manage hierarchical, nested data structures typically found in object-oriented data like CityJSON. As a result, alternative approaches have been developed to address the challenges of storing CityJSON in NoSQL databases.

Furthermore, NoSQL databases possess a flexible schema and the ability to naturally model hierarchies and complex relationships. It provides a more intuitive and efficient solution for managing the intricate structure of CityJSON data. Document and graph-oriented NoSQL databases have been explored as alternatives to CityJSON to address the limitations of relational databases in handling object-oriented data. Both MongoDB document and RDF-based graph databases have been widely utilised for this purpose. MongoDB was explored by Nys and Billen [16] and Karin et al. [22] for storing CityJSON, with evaluations comparing its performance with PostgreSQL. Nys and Billen [16] proposed a simplified schema for a document database where CityJSON components are decomposed into first-order or discriminated schemas. Their work also included a visualisation framework built using a MERN stack API architecture. Meanwhile, Karin et al. [22] focused on the querying capabilities of MongoDB to compare the API querying of CityJSON data via GraphQL against PostgreSQL. Both studies found that the querying performance of CityJSON data using MongoDB is promising. This advantage is attributed to MongoDB reduced reliance on tables and joins compared to relational database, resulting in a simpler and more efficient query of CityJSON components.

Akin and Cömert [2] developed a converter that maps CityJSON components into RDF triples by using CJIO to transform CityJSON into a DataFrame, which is then translated into RDF triples in Neo4j. While their work demonstrates the potential of graph databases for storing CityJSON, it lacks evaluation or comparison with other databases, leaving the practical effectiveness of RDF for this purpose underexplored. RDF triples, structured rigidly as subject-predicate-object, are less suited for representing the object-oriented and nested nature of 3D city models. RDF also struggles with the semantic richness and hierarchical depth of 3D city models due to its lack of internal node structure. Attributes must be expressed as additional triples rather than embedded directly into nodes, resulting in a more complex and verbose graph structure. This limitation hampers the representation of semantically rich data, making RDF less ideal for dynamic urban applications [7]. In contrast, LPG allows attributes to be directly embedded in nodes and edges, offering a more flexible and expressive approach for preserving and enriching the semantics of 3D city models. Additionally, RDF requires joins between triples for deep graph traversals, which increases query complexity [20], whereas LPG supports native and efficient traversal operations, thus enhancing performance for complex queries.

LPG graphs models consist of fundamental graph elements (i.e., nodes and edges), which can be enriched with key-value properties. The ability to annotate both nodes and edges enhances the expressiveness of the graph, allowing it to capture complex, object-oriented structures more naturally. In this study, LPG graph is utilised to model and store the nested structure of CityJSON by taking advantage of its ability to annotate both nodes and edges with properties. This flexibility allows information to be stored contextually where descriptive properties of each JSON object are embedded as node attributes, while associative attributes like those linked to City Objects are captured through edge attributes. Each JSON object is represented as an individual node, enabling a clear and expressive mapping of the hierarchical and semantic relationships inherent in CityJSON data.

3 Methodology

Although RDF-based graphs have been explored for CityJSON data management, we argue that LPGs are better suited for managing 3D city model data. As 3D city models are developed to improve interoperability across diverse applications, storing them as attributed nodes and edges within an LPG structure is more appropriate since it allows better expressiveness of information compared to RDF, particularly as LPG graphs are built on a key-value pair [30]. This approach enables the seamless association of attributes relevant to various urban applications that facilitate the semantic enrichment of 3D city models.

Junxiang et al. [32] have raised concerns about the limitations of RDF triples for storing and querying building information data. Specifically, RDF graphs and their query languages lack efficiency to support graph traversals, which poses challenges for graph querying and analysis. As a result, RDF triples often must be converted into LPG graphs to enable scalable graph analytics and fully capture complex semantic relationships [30, 3, 18, 9]. Furthermore, LPG graphs simplify data integration compared to RDF graphs [31]. This makes LPGs a more effective choice for managing 3D city models that support semantic complexity and allow efficient graph traversals and analytics. Therefore, this study aims to develop a schema model for storing CityJSON in ArangoDB, a multi-model graph database that supports the LPG graph structure.

3.1 Schema for Multi-Model Graph Database

Our approach focuses on managing 3D urban data using a multi-model graph database. The structure of an LPG-based graph database is considered a more intuitive approach compared to RDF-based graph databases and relational databases for handling complex 3D urban data.

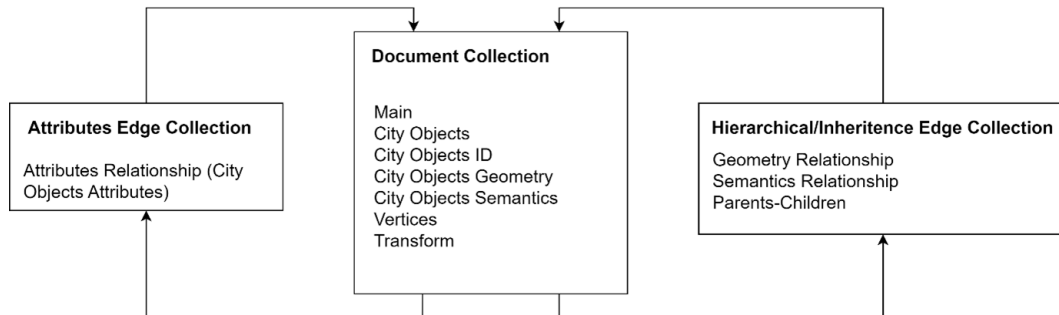
3D urban data are characterised by their complex structure, nested relationships, and semantically rich attributes. Representing such data using an object-oriented approach is more suitable, which can be achieved effectively with an LPG-based property graph. LPG structures mirror real-world object relationships more naturally, making them ideal for 3D spatial data where objects like buildings, geometry, and attributes can be abstracted through graph elements.

In ArangoDB, records are stored in JSON format as documents in document collections, while relationships between records are established and stored in edge collections where each edge references the unique key of the connected records. This design essentially treats each record in the document collection as a node, whereas the edges can be connected between nodes to represent their relationships, hence the multi-model nature of ArangoDB. Storing an entire CityJSON document as a single entity is possible in ArangoDB; however, querying and retrieving specific information can become complex due to the deeply nested structure of CityJSON components. Filtering data requires the query process to traverse the deep hierarchical levels of the CityJSON document, which can be inefficient and time-consuming. To address this issue, it is necessary to decompose CityJSON files into distinct components and store them as separate documents within a document collection.

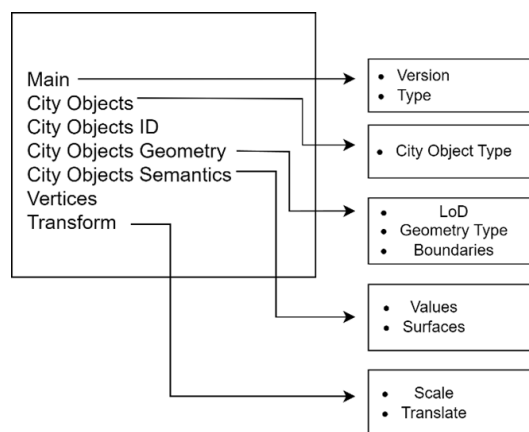
Therefore, we propose a schema for multi-model graph database to store CityJSON based on three collections, one document collection, and two edge collections (see Figure 1). The document collection gathers all the decomposed CityJSON components as individual documents. First-level objects, such as City Objects, are stored as separate documents, while second-level objects, including geometries nested within City Objects, and third-level objects, like the semantics nested within geometry, are further decomposed and stored as individual

2:6 Multi-Model Graph for CityJSON Management

documents. Additionally, each City Object ID is stored as a document to explicitly associate City Objects with their attributes and facilitate parent-child relationships. The content of each decomposed document is shown in Figure 2.



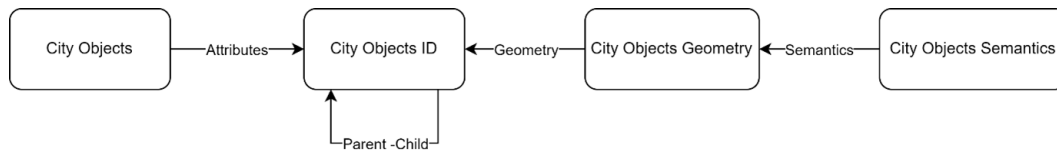
■ **Figure 1** CityJSON Multi-Model Graph Database Schema.



■ **Figure 2** Components of Decomposed CityJSON Documents.

Two edge collections are used to model relationships. The first edge collection links City Object documents to their City Object IDs and stores attributes as edge attributes. Attributes originally stored within the “Attributes” key of each City Object in the original CityJSON file are separated and mapped as key-value pairs in the edge collection. This separation of attributes aims to avoid the attributes being kept nested under their City Objects. It ensures that attributes can be queried more efficiently, and new attributes can be added or modified using the basic insert and update database operations. In the case where the City Objects do not contain any attributes, the relationship between the City Objects document and each City Object ID will still be established. This ensures that any future attributes relevant to any City Objects can be inserted.

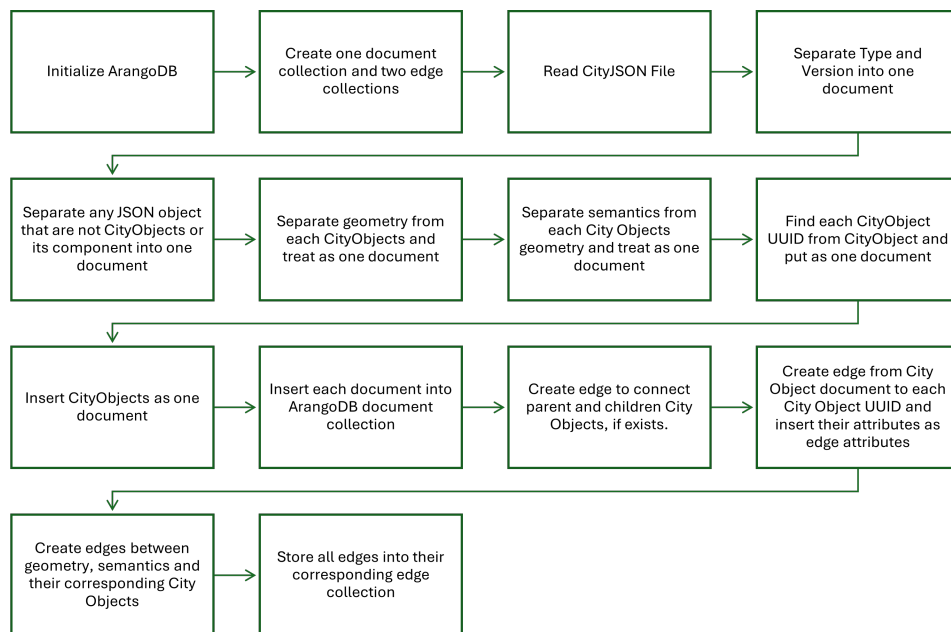
The second edge collection models the relationships between geometry, semantics, and their corresponding City Objects alongside the parent-child hierarchies among City Objects. Figure 3 illustrates the representation of the relationship between the decomposed components of CityJSON.



■ **Figure 3** Relationship Representation between Decomposed CityJSON Components.

3.2 Storing CityJSON into Multi-Model Database

A program is developed to store a CityJSON file into ArangoDB based on the schema explained in Section 3.1. The programme workflow is shown in Figure 4.



■ **Figure 4** CityJSON to LPG Graph Workflow.

The integration of CityJSON into the graph database process begins by setting up the ArangoDB environment, a graph database to store and manage CityJSON data. First, the ArangoDB graph is initialised and a dedicated database is created. Within this database, a document collection is established to store the CityJSON components. Edge collections are created to represent the relationships between these components. One edge collection is for connecting CityJSON components with their parent-child relationship and other inheritance relationship, while another edge collection is created to store CityObject attributes.

The workflow then reads the CityJSON file where the content will be decomposed into individual components with smaller and manageable JSON scripts containing the decomposed JSON objects. The code will analyse the file to identify and separate each CityJSON component with the exception of CityObjects and its components, which will be processed at a later stage. The type and version components are stored as one document (Main document, see Figure 1), which is named based on the file name of the input CityJSON file. Each of these identified components is then inserted as individual documents into the document collection, which has been created at the database initialisation step.

Next, the workflow proceeds to parse the City Objects and their components. The City Object components will be analysed and decomposed based on the following considerations:

1. Each CityObject is examined to determine its individual ID (CityObjectUUID). For each City Object ID, a document is created and inserted into the document collection.
2. The hierarchical structure of CityObjects is addressed by examining the parent-child relationships among CityObjects. If such relationships exist, edges are created to represent this relationship.
3. If a CityObject contains geometry information, the geometry is separated and stored in a dedicated document and later kept in the document collection. Edges are created to link the geometry data to its respective CityObjects.
4. If the geometry of the CityObject contains semantic information, the semantic is separated and stored in a dedicated document and kept in the document collection. Edges are created to link the semantic information to its respective geometry documents.
5. Edges are created between CityObject and all CityObject UUID and stored in the attribute edge collection regardless of whether the CityObject has attributes or not. If the CityObject contains attributes, it is inserted as edge attributes; otherwise, the edge will act as a placeholder for future attributes.

At the end of the workflow, all CityJSON components (objects, semantics, and geometry) and their attributes are stored as structured documents in the database. Relationships between these components are encoded as edges, making it possible to query and analyse the data using graph-based operations. The attributes of City Objects are stored inside a dedicated collection, which allows users to dynamically include any information regarding the CityObjects pertinent to any applications in the future. The integration process transforms the CityJSON data into a form that is highly suitable for advanced applications like urban data management and semantic querying, thus transforming urban management decision-making towards knowledge-driven initiatives.

4 Results and Analysis

4.1 CityJSON as Graph

A graph-based representation of the CityJSON data can be constructed by adhering to the workflow for storing CityJSON data in ArangoDB as outlined in Section 3.2 and structured according to the schema described in Section 3.1. For implementation and evaluation, we use three tiles of CityJSON data retrieved from the 3DBAG website ². The dataset includes multiple LoD, resulting in multiple geometries and corresponding semantics for each LoD.

The graph in Figure 5 shows the CityJSON data structure based on the components and relationships outlined in Figure 3 where the City Object documents serve as central nodes and all City Objects ID converge. It visualises how the decomposed components are interconnected and captures the hierarchical parent-child relationships among the City Objects as well as the inheritance of semantics and geometry back to their corresponding City Objects. The nodes represent the main JSON objects in CityJSON. Other information like LoD is represented as queryable attributes inside the nodes. Additionally, the graph is capable of illustrating the multiple geometries and semantics associated with each LoD, providing a comprehensive view of the structure of the original CityJSON dataset and its relationships.

² <https://3dbag.nl>

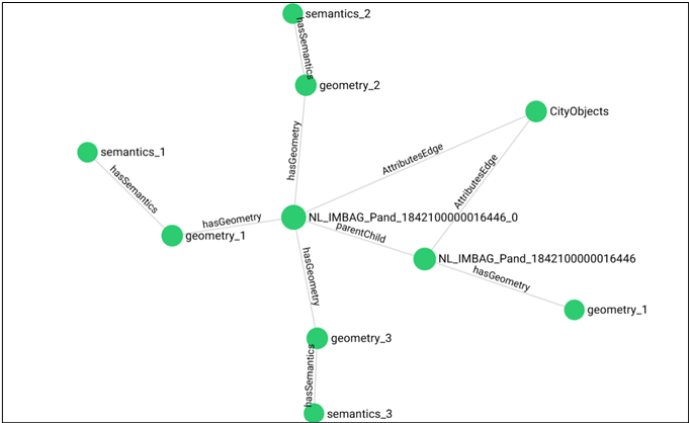


Figure 5 Representation of CityJSON Structure as Graph.

4.2 Evaluation against Relational Database

Our approach is evaluated against PostgreSQL based on the CJDB schema [17]. It uses three tables to store the CityJSON components and their relationship. The first table (`city_object`) stores the City Objects and the information describing the City Objects. The second table (`cj_metadata`) stores the information of the imported CityJSON file. Finally, the third table (`city_object_relationships`) stores the relationship between City Object, such as the parent-child relationship. The CJDB schema is shown in Figure 6.

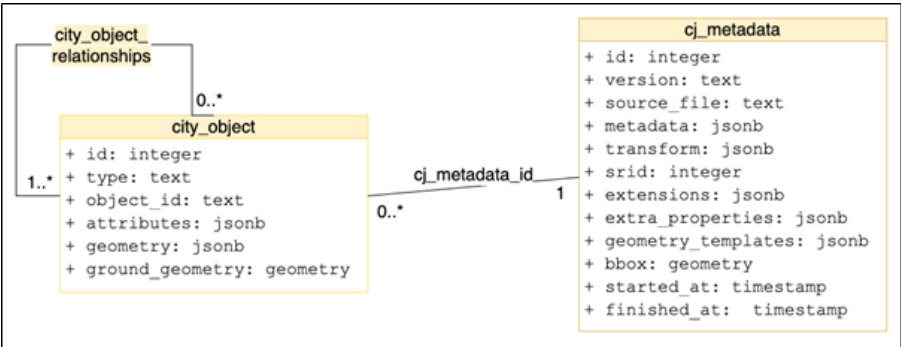


Figure 6 CJDB Schema.

As ArangoDB is a multi-model database, we evaluate our approach according to document-based query, graph-based query, and hybrid query (combination of document-based and graph-based query). It involves using three tiles from the CityJSON dataset retrieved from the 3DBAG website. Table 1 contains a description of the CityJSON datasets used for the implementation and evaluation of our approach, while Figure 7 illustrates the visualisation of the datasets using Ninja CityJSON viewer ³ [25].

The data is a multiple LoD data based on the improved LoD specification by [6]. Table 2 explains the query and the query type for the evaluation of our approach.

³ <https://ninja.cityjson.org/>

2:10 Multi-Model Graph for CityJSON Management

■ **Table 1** CityJSON Datasets Retrieved from 3DBAG Website.

Dataset	File Size (Kb)	Number of City Objects	LoDs
F-8-264-552	1409	396	LoD0, LoD1.2, LoD2.2, LoD2.3
E-10-278-556	2954	892	LoD0, LoD1.2, LoD2.2, LoD2.3
G-8-328-528	19868	6966	LoD0, LoD1.2, LoD2.2, LoD2.3

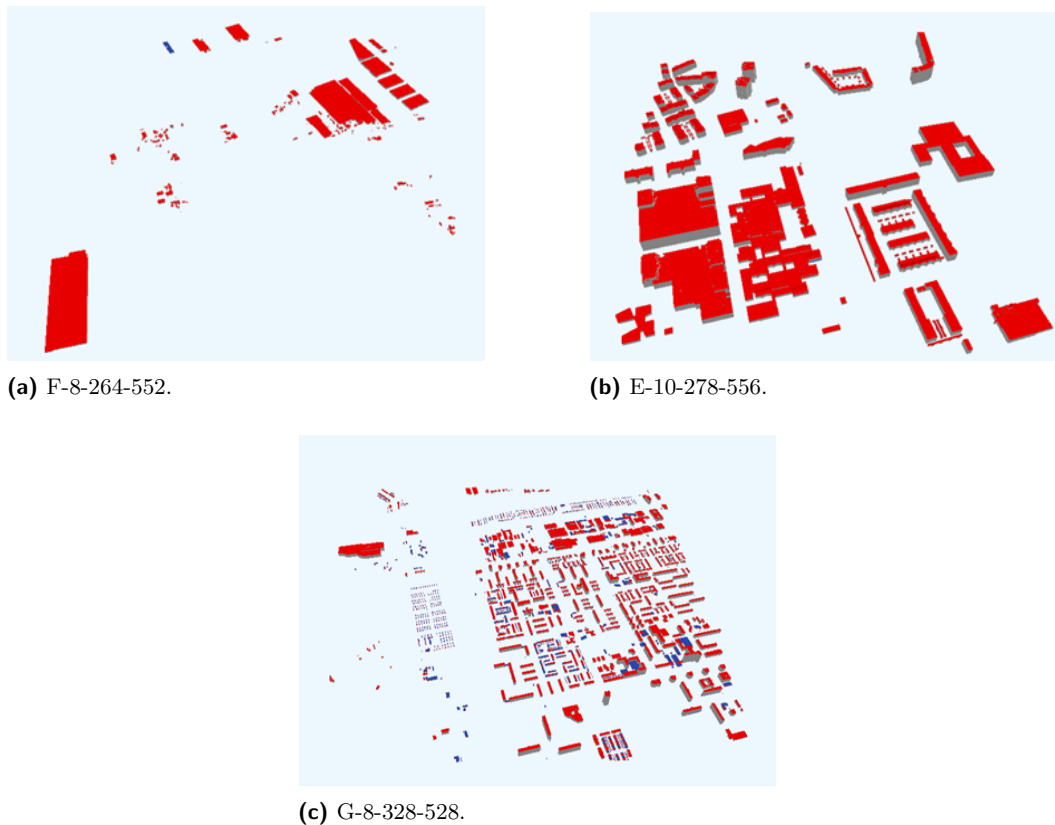
■ **Table 2** Queries for Benchmarking with PostgreSQL.

No.	Query	Query Type
Q1	Query all City Objects with “Building” type	Document
Q2	Query all LoD 1.2 City Objects	Document
Q3	Query all City Objects with slanted roof	Graph
Q4	Query City Objects with specific child	Graph
Q5	Insert “owner” attributes for all City Objects with “Building” type	Hybrid
Q6	Delete “owner” attributes for all City Objects with “Building” type	Hybrid

All benchmarks were conducted on a machine running on AMD Ryzen 5 CPU, 16 GB RAM, and a 516 GB SSD. ArangoDB 10.1 and PostgreSQL v17 with PostGIS extension were used. Each query was executed three times and the average execution time was recorded. All benchmarks were performed under warm cache conditions.

Query 1 and Query 2 are document-based query. The City Objects’ type is stored within City Object documents in the document collection, whereas the LoD is stored in geometry document. In Q3, graph-based querying is employed to identify City Objects with slanted roofs by querying the edge attributes between City Objects document and its City Object ID. Similarly, Q4 leverages graph traversal to retrieve buildings with specific child elements by querying the hierarchical relationship between City Object ID. Q5 and Q6 are document and graph queries, respectively. The City Objects type is stored in the City Objects document, while the attributes are stored as edge in the attributes edge collection. Therefore, both Q5 and Q6 must navigate the elements in the document and graph structures to complete the query. The query performance comparison is visualised in Figure 8.

The evaluation shows that our multi-model schema excels in most evaluation cases than PostgreSQL based on the CJDB schema. ArangoDB stores information natively in JSON format, while CJDB stores CityJSON information in JSONB format. Although both are stored in document-based format, ArangoDB, which is purposely built as a multi-model database that natively supports document-based data, is inherently more efficient for querying such data compared to PostgreSQL, which relies on its table-based schema to manage JSONB. This advantage is evident in Q1 and Q2 where ArangoDB achieves significantly lower execution times for querying document-based data. Additionally, ArangoDB stores information as a single document, while PostgreSQL may rely on The Oversized-Attribute Storage Technique (TOAST) table to store oversized data. This will introduce additional overhead as join operation with the TOAST table is needed when accessing oversized documents. The design advantage enables ArangoDB to store and retrieve large CityJSON documents more efficiently, making it better suited for querying large datasets.

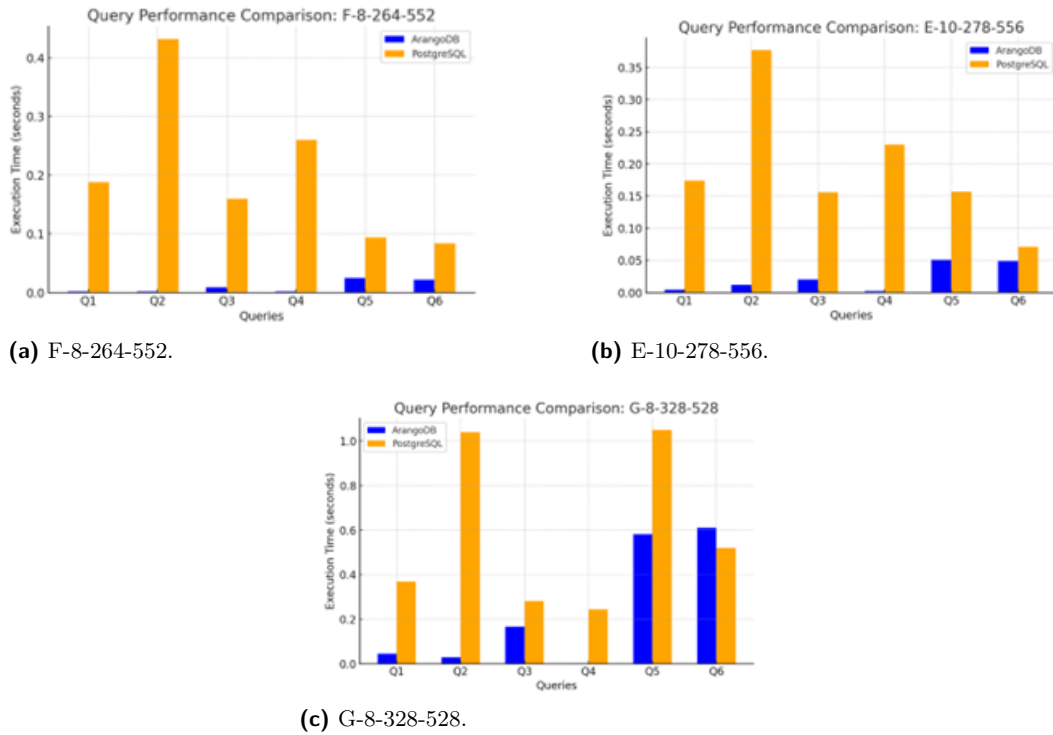


■ **Figure 7** Visualization of each Evaluation Dataset.

Q3 in ArangoDB is a graph traversal operation to retrieve City Objects attributes stored as edge attributes. The edge points towards City Object ID, which is the object that the attributes belong to. Edge attributes that are relevant to LPG make filtering more efficient since the query interacts directly with the graph structure. On the other hand, relational database relies on parsing the JSONB column and filtering its key-value pair based on JSONB operators. This approach requires a scan or index-based lookup of the JSONB column, which is computationally more expensive than edge filtering in ArangoDB. When compared to RDF, they do not natively support attributes on edges and require workarounds like reification, which will further complicating queries [30].

Meanwhile, Q4 is also a graph traversal operation to navigate the parent-child relationships between City Objects. ArangoDB handles this straightforwardly by establishing a relationship between a parent and their children. Meanwhile, PostgreSQL handles this by establishing a new table to join parent and children. Therefore, the join operation is unavoidable for PostgreSQL to query parent-child relationships, which usually will result in additional computational overhead than graph traversals. This is evident by the execution time shown in Table 3. Compared to RDF-graph, LPG-based graph databases are better suited for this purpose than RDF triples as the cost of graph traversals in RDF is higher compared to LPG [3]. RDF may require joining of multiple triples to accomplish deep graph traversal operations. Therefore, traversing the deep hierarchical relationship between parent and child is more efficient using LPG.

2:12 Multi-Model Graph for CityJSON Management



■ **Figure 8** Query Performance Comparison for each Dataset.

Q5 and Q6 in ArangoDB are hybrid queries because they involve accessing data stored in both the document collection and the edge collections. The City Object types are filtered in the document collection, while the attributes are stored as edge attributes in the graph structure. This dual-querying process introduces overhead following the need to navigate multiple data structures and perform cross-collection traversals. The overhead is particularly noticeable in Q6 (Delete operation) for the G-8-328-528 dataset where the deletion process required slightly more time than PostgreSQL. This can be attributed to the larger dataset size, which increases the complexity of traversing and modifying edge attributes after initial document filtering. While ArangoDB handles graph traversals efficiently, the combination of document filtering and edge modification introduces additional steps that slightly affect performance in large datasets. In contrast, PostgreSQL manages the deletion operation more efficiently in this specific case due to optimised JSONB operations for direct attribute modification. However, performance advantage in Q6 is limited to this specific scenario as the overall querying process still suffers from complex joins and schema rigidity in other query types. Future work could explore strategies to optimise hybrid queries in ArangoDB, such as pre-indexing relationships or implementing batch processing techniques to reduce the traversal overhead in large datasets. Despite the overhead observed in hybrid queries, ArangoDB maintains superior performance in most cases, particularly in queries involving complex relationships and semantic enrichment.

5 Conclusion and Future Works

This study adopts an object-oriented approach to abstract urban components and their relationships as graph elements using a multi-model graph database. CityJSON is decomposed into individual JSON scripts, which are stored as document nodes and linked via unique keys. Two edge collections are used: one connects City Object documents to their IDs for attribute storage while the other captures parent-child and semantic-geometry relationships. Queries are executed using document-based, graph-based, and hybrid approaches, which show better performance compared to PostgreSQL based on the CJDB schema. This demonstrates the scalability and flexibility of the proposed method.

The strength of our approach lies in the reusability of City Object nodes. Enrichment of attributes can be achieved by modifying or updating the edge attributes relevant to the City Objects to allow better expressivity. Furthermore, relationships can be established without necessitating node duplication owing to the node reusability of City Objects. This requires no joins of tables and allows a better query through graph operations, which have been demonstrated to be more time-efficient compared to query on relational model.

Representing 3D city models through object-oriented abstraction simplifies their complexity by reducing them into manageable, modular structures. Each component is treated independently, allowing flexible storage, updating, and querying. Semantic enrichment is supported by attaching attributes to nodes and edges, while topological relationships can be modelled via edges, thus enabling spatial queries [33], [23]. Future work can extend this approach to support spatial queries, including bounding box operations essential for location-based urban applications. This involves computing bounding boxes for all City Objects and storing them in the database, further enhancing spatial query capabilities for 3D city models.


References

- 1 Amgad Agoub, Felix Kunde, and MARTIN Kada. Potential of graph databases in representing and enriching standardized geodata. *Tagungsband der*, 36:208–216, 2016. URL: https://www.researchgate.net/profile/Felix_Kunde/publication/305701542_Potential_of_Graph_Databases_in_Representing_and_Enriching_Standardized_Geodata/links/579a93ea08ae2e0b31b1591a/Potential-of-Graph-Databases-in-Representing-and-Enriching-Standardized-G.
- 2 A T Akin and Ç Cömert. "CITYJSON2RDF" A Converter for Producing 3D City Knowledge Graphs. In Isikdag U. and Bayram B., editors, *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, volume 48, pages 15–20, KTU, Engineering Faculty, Trabzon, 61080, Turkey, 2024. International Society for Photogrammetry and Remote Sensing. doi:10.5194/isprs-archives-XLVIII-4-W9-2024-15-2024.
- 3 Alex Johannes Albertus Donkers, Dujuan Yang, and Nico Baken. Linked data for smart homes: Comparing RDF and labeled property graphs. *CEUR Workshop Proceedings*, 2636:23–36, 2020. URL: <https://ceur-ws.org/Vol-2636/02paper.pdf>.
- 4 Suhaibah Azri, Francois Anton, Uznir Ujang, Darka Mioc, and Alias A Rahman. *Crisp Clustering Algorithm for 3D Geospatial Vector Data Quantization*, pages 71–85. Springer International Publishing, Cham, 2015. doi:10.1007/978-3-319-12181-9_5.
- 5 Suhaibah Azri., Uznir Ujang., F. Anton, D. Mioc, and A. A. Rahman. 3D nearest neighbour search using a clustered hierarchical tree structure. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, 41(July):87–93, 2016. doi:10.5194/isprsarchives-XLI-B2-87-2016.

- 6 Filip Biljecki, Hugo Ledoux, and Jantien Stoter. An improved LOD specification for 3D building models. *Computers, Environment and Urban Systems*, 59:25–37, 2016. doi:10.1016/j.compenvurbsys.2016.04.005.
- 7 Valeriy Chernenkiy, Yuriy Gapanyuk, Anatoly Nardid, Maria Skvortsova, Anton Gushcha, Yuriy Fedorenko, and Richard Picking. Using the metagraph approach for addressing rdf knowledge representation limitations. In *2017 Internet technologies and applications (ITA)*, pages 47–52. IEEE, 2017.
- 8 Linfang Ding, Guohui Xiao, Albulen Pano, Mattia Fumagalli, Dongsheng Chen, Yu Feng, Diego Calvanese, Hongchao Fan, and Liqiu Meng. Integrating 3D city data through knowledge graphs. *Geo-Spatial Information Science*, pages 1–31, 2024. doi:10.1080/10095020.2024.2337360.
- 9 A. E.Hadi Hor, G. Sohn, P. Claudio, M. Jadidi, and A. Afnan. A semantic graph database for BIM-GIS integrated information model for an intelligent urban mobility web application. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4(4):89–96, 2018. doi:10.5194/isprs-annals-IV-4-89-2018.
- 10 Mohamad Yusoff Izham, Ujang Muhamad Uznir, Abdul Rahman Alias, Katimon Ayob, and Ismail Wan Ruslan. Influence of georeference for saturated excess overland flow modelling using 3D volumetric soft geo-objects. *Computers and Geosciences*, 37(4):598–609, 2011. doi:10.1016/j.cageo.2010.05.013.
- 11 Noraidah Keling, Izham Mohamad Yusoff, Habibah Lateh, and Uznir Ujang. *Highly Efficient Computer Oriented Octree Data Structure and Neighbours Search in 3D GIS*, pages 285–303. Springer International Publishing, Cham, 2017. doi:10.1007/978-3-319-25691-7_16.
- 12 Hugo Ledoux, Ken Arroyo Otori, Kavisha Kumar, Balázs Dukai, Anna Labetski, and Stelios Vitalis. CityJSON: a compact and easy-to-use encoding of the CityGML data model. *Open Geospatial Data, Software and Standards*, 4(1), 2019. doi:10.1186/s40965-019-0064-0.
- 13 Xiaoi Li. CityREST: CityJSON in A Database + RESTful Access, 2021.
- 14 Zulaikha Hana Mohd, Uznir Ujang, and Tan Liat Choon. Heritage house maintenance using 3D city model application domain extension approach. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, 42(4W6):73–76, 2017. doi:10.5194/isprs-archives-XLII-4-W6-73-2017.
- 15 Billy Montolalu, Siti Rochimah, and Daniel Siahaan. Sql and nosql object-database mapping using property graphs in relational cases. In *2024 11th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, pages 515–520. IEEE, 2024.
- 16 Gilles Antoine Nys and Roland Billen. From consistency to flexibility: A simplified database schema for the management of CityJSON 3D city models. *Transactions in GIS*, 25(6):3048–3066, 2021. doi:10.1111/tgis.12807.
- 17 Leon Powalka, Chris Poon, Yitong Xia, Siebren Meines, Lan Yan, Yuduan Cai, Gina Stavropoulou, Balázs Dukai, and Hugo Ledoux. cjdb: A Simple, Fast, and Lean Database Solution for the CityGML Data Model. *Lecture Notes in Geoinformation and Cartography*, pages 781–796, 2024. doi:10.1007/978-3-031-43699-4_47.
- 18 Sumit Purohit, Nhuy Van, and George Chin. Semantic Property Graph for Scalable Knowledge Graph Analytics. *Proceedings - 2021 IEEE International Conference on Big Data, Big Data 2021*, pages 2672–2677, 2021. doi:10.1109/BigData52589.2021.9671547.
- 19 Nurfairunnajiha Ridzuan, Uznir Ujang, Suhaibah Azri, and Tan Liat Choon. Visualising urban air quality using AERMOD, CALPUFF and CFD models: A critical review. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, 44(4/W3):355–363, 2020. doi:10.5194/isprs-archives-XLIV-4-W3-2020-355-2020.
- 20 Awmy Sayed and Amal Almaqrashi. Scalable and Efficient Self-Join Processing technique in RDF data. *arXiv preprint arXiv:1409.4507*, 11:43–50, 2014. arXiv:1409.4507.
- 21 Wenzhong Shi, Bisheng Yang, and Qingquan Li. An object-oriented data model for complex objects in three-dimensional geographical information systems. *International Journal of Geographical Information Science*, 17(5):411–430, 2003. doi:10.1080/1365881031000086974.

- 22 Karin Staring, Stelios Vitalis, Linda Brink, and Balazs Dukai. Combination of cityjson with postgresql, mongodb and graphql. Master's thesis, Delft University of Technology, 2020.
- 23 Muhammad Syafiq, Suhaibah Azri, and Uznir Ujang. Navigating Immovable Assets: A Graph-Based Spatio-Temporal Data Model for Effective Information Management. *ISPRS International Journal of Geo-Information*, 13(9):313, 2024. doi:10.3390/ijgi13090313.
- 24 Uznir Ujang, Francesc Anton Castro, and Suhaibah Azri. Abstract topological data structure for 3D spatial objects. *ISPRS International Journal of Geo-Information*, 8(3), 2019. doi:10.3390/ijgi8030102.
- 25 S. Vitalis, A. Labetski, F. Boersma, F. Dahle, X. Li, K. Arroyo Otori, H. Ledoux, and J. Stoter. Cityjson + Web = Ninja. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, VI-4/W1-20(September):167–173, 2020. doi:10.5194/isprs-annals-vi-4-w1-2020-167-2020.
- 26 Jochen Wendel, Alexander Simons, Alexandru Nichersu, and Syed Monjur Murshed. Rapid development of semantic 3D city models for urban energy analysis based on free and open data sources and software. *Proceedings of the 3rd ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics, UrbanGIS 2017*, 2017-Janua, 2017. doi:10.1145/3152178.3152193.
- 27 Nevil Wickramathilaka, Uznir Ujang, Suhaibah Azri, and Tan Liat Choon. Influence of Urban Green Spaces on Road Traffic Noise Levels: - a Review. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, 48(4/W3-2022):195–201, 2022. doi:10.5194/isprs-archives-XLVIII-4-W3-2022-195-2022.
- 28 Bruno Willenborg, Maximilian Sindram, and Thomas H. Kolbe. Applications of 3D City Models for a Better Understanding of the Built Environment. In Martin Behnisch and Gotthard Meinel, editors, *Trends in spatial analysis and modelling: decision-support and planning strategies*, Geotechnologies and the Environment, pages 167–191. Springer International Publishing, Cham, 2018. doi:10.1007/978-3-319-52522-8.
- 29 Zhihang Yao, Claus Nagel, Felix Kunde, György Hudra, Philipp Willkomm, Andreas Donaubauer, Thomas Adolphi, and Thomas H. Kolbe. 3DCityDB - a 3D geodatabase solution for the management, analysis, and visualization of semantic 3D city models based on CityGML. *Open Geospatial Data, Software and Standards*, 3(1), 2018. doi:10.1186/s40965-018-0046-7.
- 30 Zhanfang Zhao, Sung Kook Han, and Ju Ri Kim. LPG representation of the reification of RDF. *International Journal of Engineering and Technology(UAE)*, 7(3.34 Special Issue 34):562–566, 2018. doi:10.14419/ijet.v7i3.34.19382.
- 31 Junxiang Zhu, Heap Yih Chong, Hongwei Zhao, Jeremy Wu, Yi Tan, and Honglei Xu. The Application of Graph in BIM/GIS Integration. *Buildings*, 12(12), 2022. doi:10.3390/buildings12122162.
- 32 Junxiang Zhu, Peng Wu, and Xiang Lei. IFC-graph for facilitating building information access and query. *Automation in Construction*, 148, 2023. doi:10.1016/j.autcon.2023.104778.
- 33 Siyka Zlatanova, Alias Abdul Rahman, and Wenzhong Shi. Topological models and frameworks for 3D spatial objects. *Computers and Geosciences*, 30(4):419–428, 2004. doi:10.1016/j.cageo.2003.06.004.

Enriching Location Representation with Detailed Semantic Information

Junyuan Liu ✉ 

SpaceTimeLab, University College London, UK

Xinglei Wang ✉ 

SpaceTimeLab, University College London, UK

Tao Cheng¹ ✉ 

SpaceTimeLab, University College London, UK

Abstract

Spatial representations that capture both structural and semantic characteristics of urban environments are essential for urban modeling. Traditional spatial embeddings often prioritize spatial proximity while underutilizing fine-grained contextual information from places. To address this limitation, we introduce **CaLLiPer+**, an extension of the CaLLiPer model that systematically integrates Point-of-Interest (POI) names alongside categorical labels within a multimodal contrastive learning framework. We evaluate its effectiveness on two downstream tasks – land use classification and socioeconomic status distribution mapping – demonstrating consistent performance gains of 4% to 11% over baseline methods. Additionally, we show that incorporating POI names enhances location retrieval, enabling models to capture complex urban concepts with greater precision. Ablation studies further reveal the complementary role of POI names and the advantages of leveraging pretrained text encoders for spatial representations. Overall, our findings highlight the potential of integrating fine-grained semantic attributes and multimodal learning techniques to advance the development of urban foundation models.

2012 ACM Subject Classification Information systems → Geographic information systems; Computing methodologies → Knowledge representation and reasoning

Keywords and phrases Location Embedding, Contrastive Learning, Pretrained Model

Digital Object Identifier 10.4230/LIPIcs.GIScience.2025.3

1 Introduction

Spatial representations form the backbone of urban analysis, serving as essential tools for understanding and modeling complex urban systems. They underpin various applications, including urban functional distribution mapping [9, 10], land use classification [12], socioeconomic indicator estimation [11], future visitor prediction [5], and next-location prediction [8]. Traditional approaches typically encode locations as numeric coordinates or rely on spatial proximity [14, 15, 30], effectively capturing physical distance and structure. However, they often fail to capture the intricate functional interdependencies between places that drive urban dynamics.

In contrast, “patial” concepts emphasize the additional layers of meaning that humans ascribe to spaces, interpreting them through social, cultural, and functional attributes [7]. Point-of-Interest (POI) data offers a practical entry point for these attributes, as it couples spatial coordinates with descriptive names and labels. Such semantic information elucidates how different places function and interact within the broader urban landscape. Nevertheless, many existing embedding methods continue to emphasize spatial distance or simple categorical labels [9, 10, 28, 30, 31], underutilizing POI data’s finer-grained insights.

¹ Corresponding author



© Junyuan Liu, Xinglei Wang, and Tao Cheng;
licensed under Creative Commons License CC-BY 4.0

13th International Conference on Geographic Information Science (GIScience 2025).

Editors: Katarzyna Sila-Nowicka, Antoni Moore, David O’Sullivan, Benjamin Adams, and Mark Gahegan;
Article No. 3; pp. 3:1–3:15



Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Recent innovations in deep learning and natural language processing [17, 19, 4] facilitate richer semantic alignments within spatial data. Notably, multimodal contrastive learning [22] has proven effective in aligning geographic coordinates with textual descriptions, thereby enhancing the semantic depth of spatial embeddings. A prime example is CaLLiPer [27], which aligns POI types with spatial coordinates to yield improvements in downstream tasks. However, CaLLiPer treats POI types as broad categorical labels, potentially overlooking the granular detail contained in POI names. Such names often provide specific and context-rich information, ranging from “Starbucks Coffee” to “John’s Hardware Store,” which can further enrich location understanding and distinguish unique POIs. Yet, the systematic integration of POI names into general-purpose spatial embeddings through multimodal contrastive learning remains underexplored. Addressing this gap is crucial for fully capturing the nuanced semantics of urban environments and advancing more comprehensive urban representation models.

To enhance spatial embeddings with richer semantic detail, we incorporate POI names alongside type labels into a multimodal contrastive learning framework. Building on the original CaLLiPer model, we propose an extended version called CaLLiPer+. We evaluate its effectiveness in two downstream tasks – Land Use Classification (LUC) and Socioeconomic Status Distribution Mapping (SDM) – as well as in an additional location retrieval task.

Our contributions are as follows:

1. We extend the CaLLiPer framework by incorporating POI names alongside type information, resulting in a unified model, CaLLiPer+ (§3).
2. We evaluate the enriched semantic representation on two downstream tasks, showing consistent performance gains of 4% to 11% over POI-type-only models (§5.1).
3. We conduct retrieval experiments to assess the model’s ability to capture urban concepts, and show that enriched semantics and advanced text encoders lead to better conceptual understanding (§5.2).
4. We demonstrate the effectiveness of contrastive learning with a pretrained encoder for location representation, and highlight the potential of the resulting embeddings for downstream applications (§6).

2 Related Work

2.1 Word Embeddings and Sentence Embeddings

The advancement of natural language processing (NLP) has led to powerful embedding techniques that transform textual data into high-dimensional vector spaces, enabling machines to better process and understand linguistic semantics. Early word embedding models such as Word2Vec [16] and GloVe [21] revolutionized NLP by capturing semantic relationships between words based on their co-occurrence in large text corpora.

Building on these foundational methods, sentence embedding models like Sentence-BERT [17] and SimCSE [6] were developed to generate dense representations of entire phrases or sentences while preserving contextual nuances. More recently, large language models (LLMs) such as BERT [3], GPT [23, 2, 20], and LLaMA [25] have further enhanced text embedding capabilities, facilitating sophisticated semantic extraction across various textual contexts, including POI descriptions and names.

These advancements in NLP offer new opportunities to incorporate linguistic semantics into geospatial models, enabling the embedding of POI names and descriptions to enrich spatial representations beyond purely numerical features.

2.2 Spatial Embeddings with POIs

Spatial embedding techniques aim to encode geographic entities into vector spaces, capturing their spatial and functional relationships. POI data, which contains both geographic coordinates and semantic attributes, has been widely utilized in urban studies for tasks such as land use classification, urban function recognition, and socioeconomic mapping.

Early approaches to spatial embeddings primarily leveraged POI categories to model urban entity co-occurrence. Yao et al. [30] proposed a method that traversed POIs within a geographic region using shortest-path algorithms to extract co-occurrence patterns. Place2Vec [28] applied a K-nearest neighbor (KNN) sampling strategy with distance decay to model spatial proximity, while Doc2Vec [18] treated urban regions as documents and POIs as words, learning region embeddings based on the co-occurrence of POI categories within predefined spatial boundaries. These methods effectively captured the functional composition of urban spaces but treated POIs as categorical variables, overlooking their individual characteristics and richer semantic meanings.

To provide more distinguishing information for individual POIs, recent methods have explored integrating additional semantic attributes into spatial embeddings. Huang et al. [9] introduced the Semantics-Preserved POI Embedding (SPPE) model, which incorporates both spatial co-occurrence patterns and categorical semantics to enhance the representation of POI distributions. Similarly, HGI [10] employed hierarchical graph-based embeddings to capture multi-level semantic relationships among POIs, urban regions, and cities. While these methods improved the semantic richness of spatial representations, they still primarily rely on categorical classifications and predefined spatial structures, limiting their adaptability to diverse urban environments.

Existing methods for spatial embeddings primarily aggregate POI information within predefined regions or construct complex spatial contexts to infer urban functions. These approaches often rely on indirect or coarse-grained representations. With the growing availability of detailed POI datasets and advances in NLP, a more direct and efficient approach is to embed individual POIs by leveraging their inherent semantic information, such as names, which provide fine-grained functional and cultural context.

2.3 Multimodal Contrastive Learning for Geospatial Data

Multimodal contrastive learning has recently gained traction as an effective method for aligning heterogeneous data sources, enabling the integration of spatial coordinates with diverse information. This approach leverages contrastive objectives to maximize similarity between aligned data pairs (e.g., a location and its textual description) while distinguishing them from unrelated samples.

UrbanCLIP [29] proposed a pre-training approach for urban region representation by generating textual descriptions for satellite images using large language models and training an image encoder via a CLIP-like framework. Similarly, GeoCLIP [26] and SatCLIP [13] extended contrastive learning to geospatial data by aligning satellite imagery with geographic coordinates, supporting tasks such as geo-localization and environmental monitoring. The CaLLiPer model [27] advanced this concept by aligning POI type semantics with spatial coordinates through multimodal contrastive learning, demonstrating improved performance in land use classification and socioeconomic status mapping.

Despite these advances, existing models primarily focus on solely POI type or complex visual data, overlooking the potential benefits of simply incorporating distinguishing semantics of POI names into contrastive learning settings, which contain rich, context-specific

information that can enhance the semantic depth of spatial embeddings, offering more nuanced insights into urban functions and structures. The underutilization of POI names in multimodal frameworks is still a significant gap in current geospatial representation learning research.

3 Methodology

3.1 Overview

This study builds upon the CaLLiPer framework [27], a multimodal contrastive learning model designed to align spatial coordinates with semantic information extracted from POI data. While the core architecture remains consistent with CaLLiPer, we introduce a key modification: the integration of POI names into the textual descriptions, enriching the semantic representation of urban spaces.

Figure 1 illustrates the overall architecture, which consists of three key components: a location encoder, a text encoder, and a projection layer. These components are jointly optimized using a contrastive learning objective to align spatial and semantic information effectively.

Location encoder. The location encoder maps spatial coordinates into a continuous vector space. It applies a positional encoding function to transform raw geographic coordinates into structured representations, followed by a fully connected neural network to generate location embeddings. In this work, we apply the Grid [14] positional encoding function.

Text encoder. The text encoder is a frozen pretrained embedding model, such as Sentence-BERT [17], LLaMA [25], or GPT [20], which generates semantic embeddings from the enriched POI descriptions. By incorporating POI names alongside categorical information, it captures more nuanced semantic details, improving the discriminative power of the embeddings.

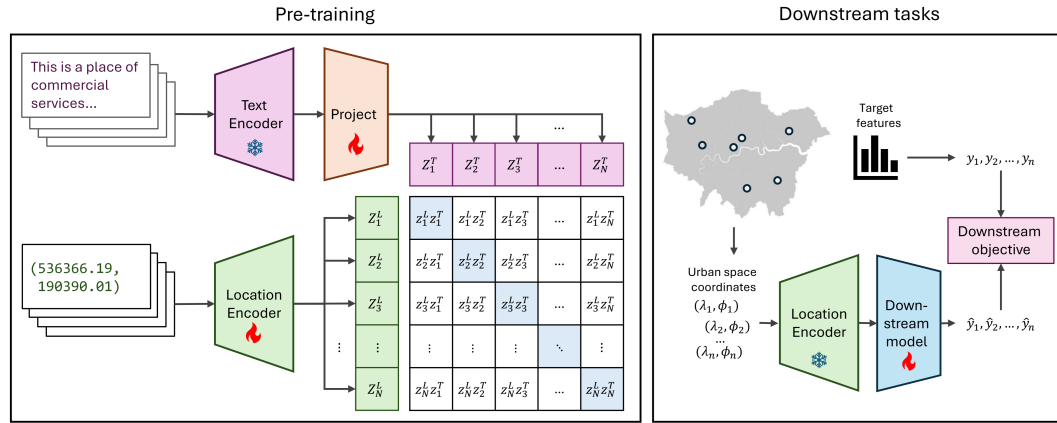
Projection layer. To facilitate direct comparison between spatial and textual embeddings, a linear projection layer maps both of them into a common vector space of dimension d . This projection ensures compatibility between modalities, enabling effective contrastive learning.

3.2 Enriching POI Descriptions with Names

In the original CaLLiPer model, POI semantics are represented solely by two levels of categorical labels from the Ordnance Survey. While effective for generalizing urban functions, this approach overlooks the rich, context-specific information embedded in POI names. Names often convey distinctive characteristics, such as cultural significance, brand identity, or specialized services, which are not captured by generic type labels. For instance, “McDonald’s” may evoke a different functional connotation compared to a generic “restaurant,” particularly in terms of cuisine style or consumption level.

To address this limitation, we extend the POI descriptions by integrating names directly into the semantic representation. For each POI p_i , we construct a combined description d_i that incorporates the name n_i , the first-level category t_{1i} , and the second-level class t_{2i} using a templated format designed to enhance the model’s understanding of the spatial context:

$$d_i = \text{Template}(n_i, t_{1i}, t_{2i}) = \text{“A place of } [t_{2i}], \text{ a type of } [t_{1i}], \text{ named } [n_i].\text{”} \quad (1)$$



■ **Figure 1** Architecture of the CaLLiPer+ model [27]. POI names are incorporated into the textual descriptions processed by the text encoder, enhancing the semantic richness of the spatial embeddings.

This enriched template ensures that the text encoder can capture both general category information and the specific nuances associated with individual POIs. By incorporating POI names, the model captures finer-grained semantic details that improve its ability to differentiate between places within the same category. This includes recognizing brand prestige (e.g., “Hilton Hotel” vs. “Budget Inn”), specific function within the same type (e.g., “The British Museum” vs. “National Gallery”), and scale or exclusivity (e.g., “local farm market” vs. “Harrods”). This richer semantic embedding enhances the model’s capacity to represent the diversity and complexity of urban environments more accurately.

3.3 Multimodal Contrastive Learning Framework

The multimodal contrastive learning framework aligns spatial coordinates with detailed textual semantics in a shared embedding space. The goal is to ensure that a POI’s spatial representation is closely aligned with its semantic description, while unrelated pairs are pushed apart.

Each POI is represented by two embeddings:

$$z_i^{(s)} = f_s(\mathbf{x}_i) \quad (\text{spatial embedding}) \quad (2)$$

$$z_i^{(p)} = W_t f_t(d_i) \quad (\text{textual embedding with name and type}) \quad (3)$$

where f_s is the spatial encoder that transforms the geographic coordinates \mathbf{x}_i into a vector representation, and f_t is a pretrained text encoder that processes the enriched POI descriptions d_i , followed by a projection layer W_t to align the dimension with spatial embedding. The inclusion of POI names in d_i ensures that the text embeddings capture both high-level categorical information and fine-grained, context-specific details.

Contrastive learning objective. The alignment between spatial and textual embeddings is achieved using the InfoNCE loss [22], which encourages positive pairs (i.e., a POI’s location and its enriched description) to be similar, while pushing apart negative pairs (i.e., mismatched locations and descriptions). The loss is defined as:

$$\mathcal{L} = -\frac{1}{2N} \left[\sum_{i=1}^N \log \frac{\exp(z_i^{(s)} \cdot z_i^{(p)} / \tau)}{\sum_{j=1}^N \exp(z_i^{(s)} \cdot z_j^{(p)} / \tau)} + \sum_{i=1}^N \log \frac{\exp(z_i^{(p)} \cdot z_i^{(s)} / \tau)}{\sum_{j=1}^N \exp(z_i^{(p)} \cdot z_j^{(s)} / \tau)} \right], \quad (4)$$

where \cdot denotes cosine similarity between embeddings, and τ is a temperature parameter that controls the sharpness of the distribution. This symmetric loss is applied to both spatial-to-textual and textual-to-spatial alignment, ensuring consistent alignment of embeddings from both modalities.

Advantages of enriched semantics. Incorporating POI names into the contrastive framework enhances the model’s ability to capture fine-grained urban semantics. The enriched descriptions provide the following benefits:

- Improved discrimination: The model can better differentiate between places within the same category by leveraging unique names.
- Context awareness: Names often imply cultural, historical, or functional context, enriching the model’s understanding of urban environments.
- Enhanced transferability: The enriched embeddings generalize better across diverse tasks.

In summary, our approach enhances the original CaLLiPer framework by incorporating POI names into the textual descriptions, leading to richer, more discriminative spatial embeddings through multimodal contrastive learning.

4 Experiments

4.1 Experimental Setup

To evaluate the impact of incorporating POI names into the spatial-semantic embeddings, we conducted experiments on two urban analytics tasks: Land Use Classification (LUC) and Socioeconomic Status Distribution Mapping (SDM). Additionally, we performed location retrieval to observe the model’s ability to capture high-level urban concepts.

4.2 Datasets

Point-of-Interest data. We use POI data from the Ordnance Survey via Digimap ², covering the Greater London area. The dataset contains approximately 340,000 POIs, each with geographic coordinates, a name, and categorical labels. POIs are classified into a hierarchical taxonomy. These data provide detailed spatial and semantic insights into London’s urban environment.

Land use data. We obtain land use data from the Verisk National Land Use Database ³, which provides high-resolution classification of land use types. The dataset includes ten primary land use categories. To create the evaluation dataset, we sample locations with a 200-meter radius buffer, ensuring balanced representation across categories.

Socioeconomic data. We obtain socioeconomic data from the Office for National Statistics (ONS) 2021 Census ⁴, specifically the National Statistics Socioeconomic Classification (NS-SeC). This dataset provides a detailed classification of socioeconomic status based on employment type, occupational hierarchy, and educational attainment. The data are aggregated at the Lower-layer Super Output Area (LSOA) level, encompassing 4,994 LSOAs across London. Each LSOA contains proportions of 1000 to 3000 residents within different occupational classes.

² <https://digimap.edina.ac.uk/>

³ <https://digimap.edina.ac.uk/roam/map/verisk>

⁴ <https://www.ons.gov.uk/>

4.3 Baselines

To assess the effectiveness of our enhanced model, CaLLiPer+, we compare it against the following baselines:

- **TF-IDF** [24]: A term frequency-inverse document frequency model that represents each region based on the POI categories within it.
- **LDA** [1]: A probabilistic topic modeling approach that infers latent topics from POI distributions, capturing urban functional structures through topic-word distributions.
- **Place2Vec** [28]: A spatial embedding model that learns representations of POIs based on their spatial co-occurrence, modeling functional similarity through a skip-gram framework.
- **Doc2Vec** [18]: A document embedding approach that treats urban regions as documents composed of POI categories, learning region representations through unsupervised learning.
- **SPPE** [9]: A semantics-preserving POI embedding method that captures spatial co-occurrence patterns and topological structures of POIs through a graph-based approach.
- **Space2Vec** [14]: A geospatial representation learning model that encodes locations through positional encoding and neural networks, learning embeddings directly from spatial coordinates.
- **CaLLiPer** [27]: The original multimodal contrastive learning model, which encodes POI categories as textual descriptions but does not incorporate POI names.

4.4 Downstream Tasks and Evaluation Metrics

We evaluate the learned spatial representations on LUC and SDM tasks. To systematically analyze the effectiveness of the learned embeddings, we employ two types of downstream models: (1) a linear model, implemented as a single-layer neural network, testing the raw expressiveness of the embeddings, and (2) a nonlinear model, implemented as a multi-layer perceptron (MLP) with a single hidden layer to capture more complex relationships.

Land use classification is a multi-class classification task that predicts the land use type of a given spatial unit based on its learned representation. We train classifiers using both a linear model and a nonlinear model and evaluate performance using:

- Precision, recall, and F1 score: These metrics are macro-averaged across classes, providing a balanced evaluation of classification performance. Higher values indicate better performance.

Socioeconomic status distribution mapping is a regression task that estimates the occupational composition of urban regions using the learned embeddings. The model predicts the proportion of residents in different socioeconomic categories at the LSOA level. We train both a linear model and a nonlinear model to compare their effectiveness. Performance is evaluated using:

- L1 distance: Measures the absolute difference between predicted and actual socioeconomic distributions.
- Chebyshev distance: Captures the maximum absolute deviation between predicted and actual distributions.
- Kullback-Leibler (KL) divergence: Evaluates the difference between the predicted and actual probability distributions, indicating how well the model captures the socioeconomic structure.

By testing the embeddings across both classification and regression tasks, and using both linear and nonlinear models, we assess their generalizability and effectiveness in capturing the information of urban environments.

4.5 Implementation Details

All models were implemented using PyTorch and trained on a machine equipped with an NVIDIA A6000 GPU. The text encoder was based on Sentence-BERT by default, which processed the enriched POI descriptions. The spatial encoder followed the same architecture as in CaLLiPer [27], using a fully connected residual network with 128-dimensional embeddings. The training process adopted a grid search approach to tune hyperparameters, resulting in a batch size of 128, a learning rate of 0.0001, and a temperature parameter of 0.07. The optimizer was Adam. The models were trained for 100 epochs with early stopping based on validation loss, and each downstream task experiment was repeated five times with different random seeds to ensure robustness. The reported results represent the mean performance across these runs.

4.6 Location Retrieval

We observe the model’s ability to retrieve urban concepts based on semantic queries. This task shows how well the learned embeddings capture urban concepts by matching textual embeddings to spatial embeddings.

Given a natural language query, we compute its embedding using a pretrained language model. We use two text encoding approaches: (1) a Sentence-Transformers model (all-MiniLM-L6-v2), which generates sentence embeddings via mean pooling over contextualized token embeddings, and (2) an OpenAI GPT-based embedding model (text-embedding-3-small), which produces a high-dimensional representation of the query and is subsequently projected into a 128-dimensional space for compatibility with the learned spatial embeddings.

The model then retrieves the most relevant locations by computing cosine similarity between the query embedding and the location embeddings of urban regions. To assess retrieval effectiveness, we visualize the top-ranked locations using geospatial maps, highlighting areas with the highest similarity to the input query.

4.7 Ablation Study

To evaluate the impact of different semantic components and text encoders, we conduct an ablation study with four model variants:

- **CaLLiPer+ GPT**: A variant that replaces the sentence transformer with GPT (text-embedding-3-small), examining the effect of a text embedding from LLM. For fairness, we only use the first 384 dimensions of the text embedding, which is the same as the default sentence transformer.
- **CaLLiPer+**: The default enhanced model that integrates both POI names and types, using a sentence transformer (all-MiniLM-L6-v2).
- **CaLLiPer+ w/o type**: A variant that removes POI types, using only POI names for textual representation.
- **CaLLiPer**: A variant that excludes POI names and relies only on POI types, which is the original CaLLiPer.

We evaluate these models on the LUC and SDM tasks. The primary metrics used are F1 score for classification and KL divergence for regression-based analysis. The results are summarized in Figure 4.

■ **Table 1** Performance comparison on the LUC task. The best and second-best performances are marked in **bold** and underlined, respectively. For better readability, all metrics are scaled by a factor of 10^2 .

Model	Linear			MLP		
	Precision \uparrow	Recall \uparrow	F1 Score \uparrow	Precision \uparrow	Recall \uparrow	F1 Score \uparrow
Random	9.6 ± 0.7	10.3 ± 0.5	9.7 ± 0.5	8.8 ± 1.3	10.3 ± 0.3	9.0 ± 0.3
TF-IDF	31.5 ± 0.4	32.2 ± 0.2	31.3 ± 0.3	31.8 ± 0.6	33.3 ± 0.5	31.7 ± 0.6
LDA	30.8 ± 0.3	29.1 ± 0.2	28.4 ± 0.2	31.5 ± 1.1	30.4 ± 0.7	29.2 ± 0.9
Place2Vec	30.9 ± 0.8	26.1 ± 0.7	26.3 ± 0.7	35.1 ± 1.2	32.7 ± 1.0	32.4 ± 1.2
Doc2Vec	32.4 ± 0.4	28.2 ± 0.1	28.0 ± 0.1	34.9 ± 0.9	33.8 ± 0.5	32.7 ± 0.6
SPPE	30.5 ± 0.4	27.0 ± 0.2	26.6 ± 0.2	34.5 ± 0.9	32.9 ± 0.7	32.2 ± 0.5
HGI	33.0 ± 0.5	30.0 ± 0.6	29.9 ± 0.6	33.6 ± 0.5	32.0 ± 0.9	31.6 ± 0.7
Space2Vec	28.6 ± 0.6	28.5 ± 0.8	27.4 ± 0.7	29.6 ± 0.6	28.9 ± 0.5	27.8 ± 0.3
CaLLiPer	36.5 ± 0.6	35.3 ± 0.2	34.6 ± 0.3	37.7 ± 0.8	35.5 ± 0.8	34.6 ± 0.8
CaLLiPer+	<u>37.5 ± 0.7</u>	<u>35.5 ± 0.5</u>	<u>35.2 ± 0.6</u>	<u>40.0 ± 0.4</u>	<u>36.0 ± 0.5</u>	<u>36.6 ± 0.5</u>
CaLLiPer+GPT	40.5 ± 0.6	36.7 ± 0.2	36.8 ± 0.3	41.3 ± 0.7	37.8 ± 0.4	37.6 ± 0.3

■ **Table 2** Performance comparison on the SDM task. The best and second-best performances are marked in **bold** and underlined, respectively. For better readability, all metrics are scaled by a factor of 10^2 .

Model	Linear			MLP		
	L1 \downarrow	Chebyshev \downarrow	KL \downarrow	L1 \downarrow	Chebyshev \downarrow	KL \downarrow
Random	30.31 ± 0.03	9.25 ± 0.01	7.73 ± 0.01	31.40 ± 0.22	9.55 ± 0.11	8.21 ± 0.14
TF-IDF	24.79 ± 0.04	7.43 ± 0.01	5.36 ± 0.01	24.36 ± 0.15	7.28 ± 0.05	5.20 ± 0.04
LDA	26.14 ± 0.01	7.84 ± 0.00	5.87 ± 0.00	25.85 ± 0.14	7.77 ± 1.12	5.80 ± 0.72
Place2Vec	23.47 ± 0.09	6.94 ± 0.02	4.81 ± 0.02	22.81 ± 0.06	6.81 ± 0.01	4.61 ± 0.02
Doc2Vec	24.01 ± 0.07	7.15 ± 0.02	4.99 ± 0.02	23.10 ± 0.19	6.89 ± 0.06	4.75 ± 0.08
SPPE	24.32 ± 0.16	7.24 ± 0.06	5.11 ± 0.06	23.63 ± 0.19	7.04 ± 0.06	4.91 ± 0.07
HGI	23.28 ± 0.08	6.93 ± 0.02	4.79 ± 0.03	22.73 ± 0.05	6.80 ± 0.02	4.60 ± 0.02
Space2Vec	25.13 ± 0.15	7.56 ± 0.04	5.65 ± 0.06	23.55 ± 0.20	7.12 ± 0.09	5.00 ± 0.08
CaLLiPer	21.63 ± 0.04	6.55 ± 0.05	4.26 ± 0.01	20.52 ± 0.14	6.24 ± 0.03	3.90 ± 0.06
CaLLiPer+	<u>20.87 ± 0.02</u>	<u>6.35 ± 0.01</u>	<u>3.98 ± 0.01</u>	<u>19.85 ± 0.19</u>	<u>6.02 ± 0.06</u>	<u>3.63 ± 0.07</u>
CaLLiPer+GPT	20.26 ± 0.03	6.09 ± 0.01	3.74 ± 0.01	19.38 ± 0.02	5.83 ± 0.04	3.47 ± 0.01

5 Results and Analysis

5.1 Performance on Downstream Tasks

Tables 1 and 2 summarize the results for LUC and SDM tasks. Across both tasks, multimodal contrastive learning models outperform traditional methods, demonstrating the effectiveness of integrating spatial and semantic information. Baseline models such as TF-IDF and LDA rely on aggregated POI type distributions within regions, limiting their ability to capture fine-grained relationships between locations. While methods like Place2Vec and Doc2Vec improve upon this by incorporating spatial co-occurrence structures, their reliance on unsupervised embedding techniques without explicit spatial-semantic alignment leads to weaker performance. In contrast, CaLLiPer and its extensions, which align POI-based textual representations with spatial coordinates, consistently achieve better results, confirming the advantages of multimodal contrastive learning.

Additionally, CaLLiPer+ achieves superior and more stable performance across all metrics. In LUC, CaLLiPer+ consistently outperforms the original CaLLiPer model, achieving higher precision, recall, and F1 scores across both linear and MLP classifiers. This demonstrates that integrating POI names alongside type-based descriptions enriches the model's semantic understanding of urban space, allowing for better land use classification. A similar trend is observed in SDM, where CaLLiPer+ further reduces errors across all three evaluation metrics, suggesting that POI names provide valuable contextual information for modeling socioeconomic distributions. Notably, CaLLiPer+ GPT achieves the best performance across both tasks, reinforcing the importance of using more powerful text encoders for spatial representation learning.

Third, the improvements observed with MLP over the linear model suggest that the learned embeddings still contain complex, non-linear relationships that can be further leveraged by downstream tasks. While baseline models such as TF-IDF and LDA show limited gains with MLP, indicating that their representations are mostly exhausted by simple classifiers, CaLLiPer-based models still exhibit a more notable performance boost. CaLLiPer+ effectively aligns spatial and semantic information, and the embeddings still retain structured patterns that require more expressive models to fully exploit, highlighting the depth and richness of the learned representations.

These findings highlight the advantages of incorporating both POI names and stronger text embedding models for geospatial representation learning, improving the model's ability to capture complex urban semantics across diverse tasks.

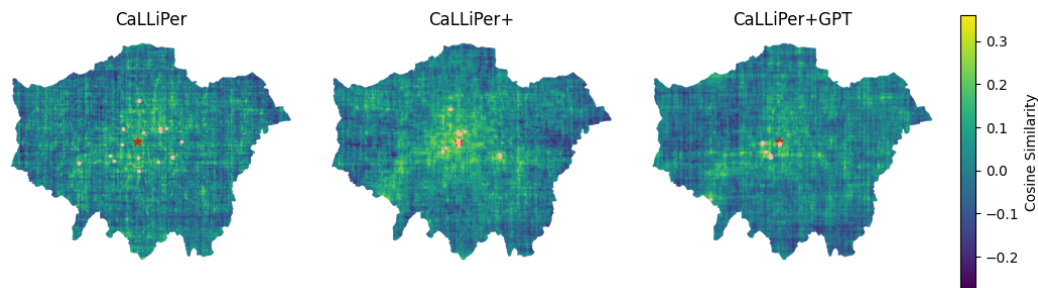
5.2 Location Retrieval

Location retrieval evaluates the model's ability to associate spatial embeddings with meaningful semantic queries, including specific place names and abstract urban concepts. The results, shown in Figures 2 and 3, illustrate how different model variants respond to retrieval tasks.

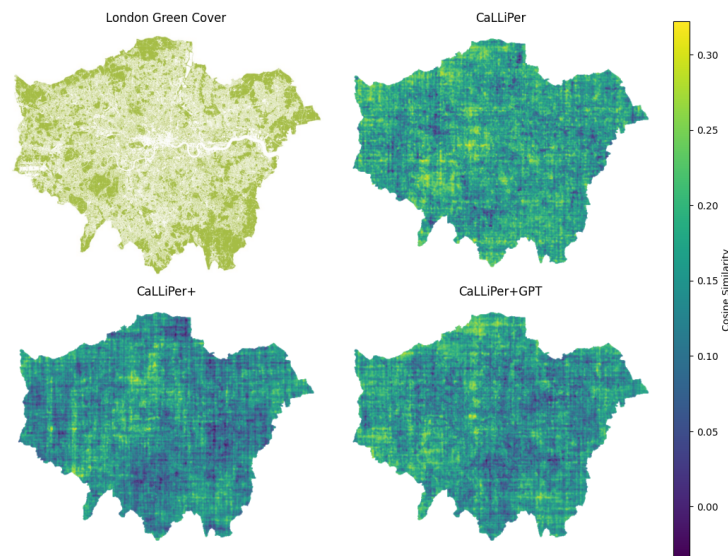
First, using POI names directly for retrieval demonstrates that including POI names in the text encoder significantly improves the model's ability to locate specific places. In Figure 2, models that incorporate POI names (CaLLiPer+ and CaLLiPer+GPT) produce more precise and concentrated retrieval results compared to the original CaLLiPer model, which relies solely on categorical types. The use of a more powerful text encoder, such as GPT embeddings in CaLLiPer+GPT, further enhances localization, leading to more accurate spatial responses.

Second, for high-level conceptual retrieval, such as identifying regions characterized by abstract urban concepts (e.g., green cover), the inclusion of POI names introduces both benefits and challenges. As seen in Figure 3, models that incorporate POI names sometimes exhibit increased dispersion in similarity scores when handling broad, high-level concepts. This suggests that when the model's semantic understanding is insufficient, in such cases, additional name-based details can introduce ambiguity. However, when equipped with a more advanced text encoder (e.g., CaLLiPer+GPT), the model can effectively utilize this additional semantic information to establish clearer distinctions between different urban functions, demonstrating improved conceptual retrieval. This improvement can be attributed to GPT's ability to capture hierarchical urban concepts and their interconnections, enabling a more nuanced understanding of spatial semantics.

Overall, our results highlight the benefits of integrating POI names in location retrieval. Name-enhanced models improve direct place retrieval and, with sufficiently strong text encoders, also facilitate better discrimination of abstract spatial concepts.



■ **Figure 2** Similarity map for “The National Gallery.” The red star is the actual location of the target, and the yellow points are the top 30 similar locations.



■ **Figure 3** Similarity map for “A place of park or green cover.” The ground truth is based on green cover data from London DataStore ⁵.

5.3 Ablation Study Results

Figure 4 presents the results of our ablation study. Both POI names and types contribute to improving downstream tasks, as seen from the superior performance of CaLLiPer+ compared to CaLLiPer and CaLLiPer+ w/o type. This suggests that combining both sources of semantic information leads to more informative spatial representations.

Interestingly, even when POI types are removed (CaLLiPer+ w/o type), the model still outperforms CaLLiPer, indicating that POI names carry richer and more discriminative semantic details than type labels alone. This highlights the potential of leveraging fine-grained textual information like POI names in spatial embedding models.

Moreover, using a stronger text encoder (CaLLiPer+ GPT) further improves results across both tasks. The enhanced semantic representation from a large language model allows for a better understanding of the text concepts in urban semantics, reinforcing the importance of high-quality embeddings in geospatial contrastive learning.

⁵ <https://apps.london.gov.uk/green-cover>

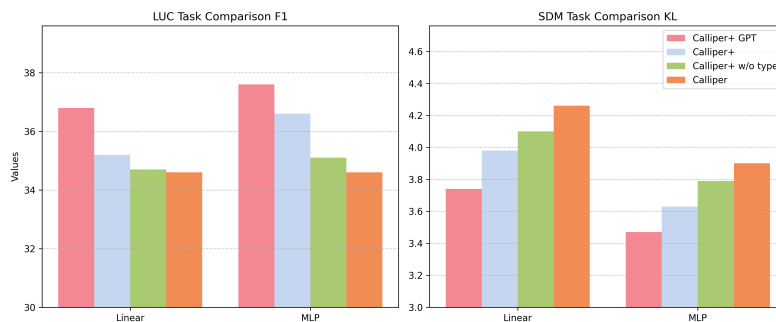


Figure 4 Ablation study results comparing model variations across LUC and SDM tasks. The left plot shows F1 score \uparrow performance on LUC, while the right plot presents KL divergence \downarrow results for SDM. All metrics are scaled by a factor of 10^2 .

6 Discussion and Conclusion

We explore the impact of integrating POI names into multimodal contrastive learning for spatial representation. By extending the CaLLiPer framework to incorporate both POI types and names, we introduce CaLLiPer+, which enhances the semantic richness of location embeddings. Our experiments across land use classification, socioeconomic status distribution mapping, and location retrieval reveal key insights into the role of enriched textual descriptions in geospatial learning.

Effectiveness of POI names in spatial representation. The combining of POI names with types in multi-modal contrastive learning improves downstream task performance consistently. POI names provide more specific and context-aware semantic signals, capturing fine-grained distinctions that categorical types alone may overlook. This effect is particularly evident in retrieval tasks, where name-enhanced models demonstrate greater precision in identifying specific locations.

Impact of text encoder strength. Using more advanced text embeddings, such as those from GPT-based models, further refines spatial representation. The CaLLiPer+ GPT model consistently outperforms others, suggesting that stronger language models contribute to a deeper understanding of urban semantics. This aligns with findings in location retrieval, where better text embeddings enable clearer conceptual differentiation, especially for high-level concepts.

Limitations and future work. The quality of spatial embeddings relies on the density and distribution of POIs across different urban areas. Regions with too sparse POI coverage may lead to less informative representations, limiting generalizability. Also, the information beyond the semantics still needs to be explored. Future work should incorporate additional modalities such as road networks, street-view imagery, and mobility patterns to enrich spatial information. Additionally, while our current downstream tasks provide initial validation, further research should explore a wider range of urban analytics applications and develop task-specific models that better leverage the structure of learned embeddings for improved adaptability and performance.

Conclusion. This work demonstrates that incorporating POI names into geospatial contrastive representation learning leads to improved performance in multiple urban analytics tasks. By aligning spatial and semantic information more effectively, CaLLiPer+ provides a more detailed and context-aware model for understanding urban environments. The effectiveness of semantic information highlights the potential of using pretrained multimodal models to generate enriched spatial embeddings in advancing urban intelligence.

References

- 1 David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003. URL: <https://jmlr.org/papers/v3/blei03a.html>.
- 2 Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- 3 Jacob Devlin, Chang Ming-Wei, Lee Kenton, and Toutanova Kristina. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, volume 1, page 2. Minneapolis, Minnesota, 2019.
- 4 Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint*, 2024. [arXiv:2407.21783](https://arxiv.org/abs/2407.21783).
- 5 Shanshan Feng, Gao Cong, Bo An, and Yeow Meng Chee. Poi2vec: Geographical latent representation for predicting future visitors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017.
- 6 Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, 2021. doi:10.18653/V1/2021.EMNLP-MAIN.552.
- 7 Michael F. Goodchild. Platial. In *International Encyclopedia of Geography*, pages 1–5. Wiley, September 2020. doi:10.1002/9781118786352.wbieg2046.
- 8 Ye Hong, Yatao Zhang, Konrad Schindler, and Martin Raubal. Context-aware multi-head self-attentional neural network model for next location prediction. *Transportation Research Part C: Emerging Technologies*, 156:104315, 2023.
- 9 Weiming Huang, Lizhen Cui, Meng Chen, Daokun Zhang, and Yao Yao. Estimating urban functional distributions with semantics preserved poi embedding. *International Journal of Geographical Information Science*, 36(10):1905–1930, 2022. doi:10.1080/13658816.2022.2040510.
- 10 Weiming Huang, Daokun Zhang, Gengchen Mai, Xu Guo, and Lizhen Cui. Learning urban region representations with pois and hierarchical graph infomax. *ISPRS Journal of Photogrammetry and Remote Sensing*, 196:134–145, 2023.
- 11 Neal Jean, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.
- 12 Neal Jean, Sherrie Wang, Anshul Samar, George Azzari, David Lobell, and Stefano Ermon. Tile2vec: Unsupervised representation learning for spatially distributed data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3967–3974, 2019. doi:10.1609/AAAI.V33I01.33013967.
- 13 Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. Satclip: Global, general-purpose location embeddings with satellite imagery. *arXiv preprint arXiv:2311.17179*, 2023. doi:10.48550/arXiv.2311.17179.

- 14 Gengchen Mai, Krzysztof Janowicz, Bo Yan, Rui Zhu, Ling Cai, and Ni Lao. Multi-scale representation learning for spatial feature distributions using grid cells. In *International Conference on Learning Representations*, 2020. URL: <https://openreview.net/forum?id=rJ1jdh4KDH>.
- 15 Gengchen Mai, Yao Xuan, Wenyun Zuo, Yutong He, Jiaming Song, Stefano Ermon, Krzysztof Janowicz, and Ni Lao. Sphere2vec: A general-purpose location representation learning over a spherical surface for large-scale geospatial predictions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 202:439–462, 2023.
- 16 Tomas Mikolov. Efficient estimation of word representations in vector space. *arXiv preprint*, 2013. [arXiv:1301.3781](https://arxiv.org/abs/1301.3781).
- 17 Reimers Nils and Gurevych Iryna. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019.
- 18 Haifeng Niu and Elisabete A Silva. Delineating urban functional use from points of interest data with neural network embedding: A case study in greater london. *Computers, Environment and Urban Systems*, 88:101651, 2021. doi:10.1016/J.COMPENVURBSYS.2021.101651.
- 19 OpenAI. Introducing chatgpt. <https://openai.com/blog/chatgpt>, 2022. Accessed: 2023-07-26.
- 20 OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. doi:10.48550/arXiv.2303.08774.
- 21 Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. doi:10.3115/V1/D14-1162.
- 22 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- 23 Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- 24 Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.
- 25 Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint*, 2023. [arXiv:2302.13971](https://arxiv.org/abs/2302.13971).
- 26 Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Advances in Neural Information Processing Systems*, 36, 2024.
- 27 Xinglei Wang, Tao Cheng, Stephen Law, Zichao Zeng, Lu Yin, and Junyuan Liu. Multi-modal contrastive learning of urban space representations from poi data. *Computers, Environment and Urban Systems*, 120:102299, 2025. doi:10.1016/J.COMPENVURBSYS.2025.102299.
- 28 Bo Yan, Krzysztof Janowicz, Gengchen Mai, and Song Gao. From itdl to place2vec: Reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts. In *Proceedings of the 25th ACM SIGSPATIAL international conference on advances in geographic information systems*, pages 1–10, 2017. doi:10.1145/3139958.3140054.
- 29 Yibo Yan, Haomin Wen, Siru Zhong, Wei Chen, Haodong Chen, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In *Proceedings of the ACM on Web Conference 2024*, pages 4006–4017, 2024. doi:10.1145/3589334.3645378.

- 30 Yao Yao, Xia Li, Xiaoping Liu, Penghua Liu, Zhaotang Liang, Jinbao Zhang, and Ke Mai. Sensing spatial distribution of urban land use by integrating points-of-interest and google word2vec model. *International Journal of Geographical Information Science*, 31(4):825–848, 2017. doi:10.1080/13658816.2016.1244608.
- 31 Wei Zhai, Xueyin Bai, Yu Shi, Yu Han, Zhong-Ren Peng, and Chaolin Gu. Beyond word2vec: An approach for urban functional region extraction and identification by combining place2vec and pois. *Computers, environment and urban systems*, 74:1–12, 2019. doi:10.1016/J.COMPENVURBSYS.2018.11.008.

Precomputed Topological Relations for Integrated Geospatial Analysis Across Knowledge Graphs

Katrina Schweikert ✉ 

School of Computing and Information Science, University of Maine, Orono, ME, USA

David K. Kedrowski ✉ 

School of Computing and Information Science, University of Maine, Orono, ME, USA

Shirly Stephen ✉ 

NCEAS, Department of Geography, University of California, Santa Barbara, CA, USA

School of Computing and Information Science, University of Maine, Orono, ME, USA

Torsten Hahmann ✉ 

School of Computing and Information Science, University of Maine, Orono, ME, USA

Abstract

Geospatial Knowledge Graphs (GeoKGs) represent a significant advancement in the integration of AI-driven geographic information, facilitating interoperable and semantically rich geospatial analytics across various domains. This paper explores the use of topologically enriched GeoKGs, built on an explicit representation of S2 Geometry alongside precomputed topological relations, for constructing efficient geospatial analysis workflows within and across knowledge graphs (KGs).

Using the SAWGraph knowledge graph as a case study focused on environmental contamination by PFAS, we demonstrate how this framework supports fundamental GIS operations – such as spatial filtering, proximity analysis, overlay operations and network analysis – in a GeoKG setting while allowing for the easy linking of these operations with one another and with semantic filters. This enables the efficient execution of complex geospatial analyses as semantically-explicit queries and enhances the usability of geospatial data across graphs. Additionally, the framework eliminates the need for explicit support for GeoSPARQL’s topological operations in the utilized graph databases and better integrates spatial knowledge into the overall semantic inference process supported by RDFS and OWL ontologies.

2012 ACM Subject Classification Computing methodologies → Spatial and physical reasoning; Computing methodologies → Ontology engineering

Keywords and phrases knowledge graph, GeoKG, spatial analysis, ontology, SPARQL, GeoSPARQL, discrete global grid system, S2 geometry, GeoAI, PFAS

Digital Object Identifier 10.4230/LIPIcs.GIScience.2025.4

Supplementary Material *Software (SPARQL queries)*: <https://github.com/SAWGraph/public/tree/main/UseCases/UC3-Tracing/UC3-CQ15/GIScience2025-queries> [52]

archived at [swh:1:dir:678ee78feb48f235c42bd5722e4c19f81f91f9dc](https://www.swh.io/dir/678ee78feb48f235c42bd5722e4c19f81f91f9dc)

Funding This research has been supported by the National Science Foundation under Grant No. 2333782: “Safe Agricultural Products and Water Graph (SAWGraph): An OKN to Monitor and Trace PFAS and Other Contaminants in the Nation’s Food and Water Systems”, which is part of a larger NSF effort to develop a Prototype Open Knowledge Network (Proto-OKN: <https://www.proto-okn.net/>). Any opinions expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Acknowledgements We are grateful to the entire SAWGraph team and all collaborators from federal and state agencies, especially Maine DEP and DACF and Antony Williams and Vasu Kilaru from EPA for their invaluable collaboration on the SAWGraph project. Furthermore, we thank the other Proto-OKN teams for their input on the spatial knowledge graph and the five anonymous reviewers for their constructive comments that helped clarify the presentation.



© Katrina Schweikert, David K. Kedrowski, Shirly Stephen, and Torsten Hahmann;
licensed under Creative Commons License CC-BY 4.0

13th International Conference on Geographic Information Science (GIScience 2025).

Editors: Katarzyna Sila-Nowicka, Antoni Moore, David O’Sullivan, Benjamin Adams, and Mark Gahegan;
Article No. 4; pp. 4:1–4:22



Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

Geospatial Knowledge Graphs (GeoKGs) represent a key advancement in AI-driven geographic information integration, enabling interoperable and semantically rich geospatial analytics across diverse domains [63, 37]. They employ a flexible linked data structure wherein data is represented as a set of interconnected entities identified by URIs that are linked to each other via relations (denoted by predicates) to form a graph of nodes and edges. Early geospatial linked datasets, such as OpenStreetMap [44] and Geonames [61], mainly focused on converting geographic data into linked data using Semantic Web standards, such as the Resource Description Framework (RDF) [47], and its semantic extensions RDFS [4] and the Web Ontology Language (OWL2) [23]. Recent GeoKGs extend this by semantically enriching the geographic data with other domain-specific and generalized knowledge to capture spatial, temporal, and thematic contexts [54]. Within GeoKGs, data (i.e. facts) and knowledge (i.e. rules that define and constrain the data schema) become interconnected. Recognizing their transformative potential to prepare data for answering many kinds of questions, several large-scale GeoKGs have been developed, including KnowWhereGraph [29], UF-OKN (Urban Flooding Open Knowledge Network) [20, 31], SAWGraph (Safe Agricultural Products and Water Graph) [19], along with many other KGs being developed under NSF’s Proto-OKN (Open Knowledge Network) [41] and its predecessor initiatives [2]. These efforts address long-standing challenges in geospatial data discovery and usability by transforming heterogeneous, cross-disciplinary geospatial datasets into FAIR (Findable, Accessible, Interoperable, and Reusable) resources [62], thus enhancing interoperability and simplifying integrated querying.

Current GeoKGs still primarily serve as semantically enriched sources of data and knowledge, whereas more advanced spatial analysis is left to traditional Geographic Information Systems (GIS) [38] or relational spatial databases [49]. However, adding explicit semantics to GeoKGs through formal ontologies [17] may allow executing many geospatial analyses directly in GeoKGs as inferential reasoning tasks. This paper explores this hypothesis by specifically focusing on how *topologically enriched* GeoKGs [56] efficiently support advanced geospatial analysis workflows within and across such graphs. To do so, we adopt and refine KnowWhereGraph’s approach [29, 56] of using an explicit representation of a discrete global grid system – S2 Geometry [50] in our case – in GeoKGs together with precomputed and materialized topological relations between geospatial entities. In our approach, here referred to as *Spatial Reference Entities with Precomputed Topological Relations* (*SRE+Topology* for short) spatial entities, such as S2 cells from S2 Geometry as well as administrative regions, serve as reference spatial entities to which geospatial features are spatially linked as a way of precomputing approximate locations and intersections.

Using the SAWGraph KGs as a case study, we demonstrate how the SRE+Topology framework can facilitate a broad range of geospatial analyses and overcome limitations of GeoSPARQL [43, 3] for querying and reasoning about spatial interactions within and across GeoKGs. In this endeavor, we concentrate on three key aspects:

1. We show how this framework supports efficient execution of fundamental GIS operations – such as spatial filtering, proximity analysis, overlay operations, and network analysis – directly in GeoKGs using existing KG technology *without the need for GeoSPARQL, specialized geospatial indexing, hybrid spatial reasoners, or explicit spatial query support*.
2. Our example queries demonstrate how the approach integrates spatial relationships into the regular semantic inference process that is facilitated by the semantics of RDFS and OWL2 in any RDF-based, semantically-enabled graph database. This deeper integration with the semantics of thematic ontologies allows easy linking of multiple geospatial operations across graphs, often within a single SPARQL query.

3. We illustrate how to perform advanced geospatial analyses by combining fundamental geospatial operations, including complementary ones such as overlay analysis and network tracing. Such integrated analyses would often become prohibitively computationally expensive in a GeoKG if relying exclusively on GeoSPARQL.

This work goes beyond the prior efforts in KWG by using the S2 grid in a GeoKG not just to facilitate a “follow-your-nose” exploration of spatially related data [56, 29] but to efficiently execute advanced geospatial analyses directly as SPARQL queries within and across GeoKGs.

2 Background & Related Work

Many GeoKGs represented using RDF, RDFS and OWL2 rely on the Open Geospatial Consortium (OGC) GeoSPARQL standard [43, 3] as vocabulary for specifying spatial geometries and constructing spatial queries. Its classes `geo:Feature` and `geo:Geometry` can describe geospatial entities and their geometries, such as points, polylines, or polygons, whose details can be encoded using WKT (Well-Known Text) strings. Furthermore, GeoSPARQL supports various geometric operations, including for distance computations (`geof:distance`), area measurements (`geof:area`), and for deriving new geometries (e.g., `geof:buffer`, `geof:intersection`, `geof:convexHull`). Additionally, it provides topological operations as both *relations between spatial objects* (i.e., predicates) and as *functions on geometries* (i.e., query functions). They include eight relations, such as `geo:sfContains`, `geo:sfOverlaps`, and `geo:sfTouches` and their functional equivalents (e.g. `geof:sfContains`), that are based on the Dimensionally Extended Nine-Intersection Model (DE-9IM) [10].

Scalability Challenges of GeoSPARQL. Most of the RDF databases that support GeoSPARQL are only partially compliant with the standard in that they only support its topological query functions but not its predicates [26, 46]. But a bigger concern is that the functions are computed dynamically at query time, which poses serious efficiency and scalability challenges [32, 16]. Even RDF databases that also implement the topological predicates, such as GraphDB¹, compute them only at query time.

Many common operations, such as arithmetic aggregations and semantic filtering, are well-optimized for SPARQL [14, 53, 58], the query language used for RDF. This is not the case for the spatial operations defined by GeoSPARQL, especially those involving spatial joins over complex geometries, which remain computationally and architecturally challenging [25, 27, 34]. This is especially true for polygon-based operations in graphs that contain high-resolution polygons or multi-polygons, which can become computationally prohibitive. The performance of such computations is influenced by various factors, including the size of the graph and the extent of federation across multiple graphs. However, one of the primary bottlenecks is that geometric computations have polynomial-time complexity relative to the number of nodes in the geometries being tested [49]. To optimize spatial querying in graph databases, various indexing techniques can be adopted, including R-tree [30], quadtree [36], and geohashing [35]. Bounding-box approximations help further reduce expensive geometric computations [8]. Hybrid architectures, such as integrations of graph and spatial databases (e.g., GraphDB + Elasticsearch), improve performance by adding specialized spatial indexing [9, 45]. Despite these optimizations, spatial operations in GeoKGs remain inefficient [39]. For example, in KnowWhereGraph [29] which contains ~29 billion statements, polygon intersection queries frequently time out. Strategies such as graph partitioning, parallel processing (GPU, Spark), caching, and distributed computation offer partial solutions but introduce significant overhead and do not fundamentally resolve the inefficiencies of query-time spatial computations.

¹ <https://graphdb.ontotext.com/documentation/10.8/geosparql-support.html>

Semantic Integration Challenges of GeoSPARQL. A second major limitation of GeoSPARQL-based GeoKGs is that when topological relations are processed at query-time, spatial querying is decoupled from the RDFS- and OWL2-facilitated semantic inferencing that graph databases afford, which prevents better integration of spatial and non-spatial knowledge. For example, while an OWL2 rule could express that “if Point A is inside region B, then contamination at B will affect A”, current graph databases do not propagate topological knowledge inferred from geometries, such as “point A is inside B”, via such semantic rules. Consequently, GeoSPARQL enables spatial queries but does not support full-fledged spatial reasoning or deeper integration with other, non-spatial semantic reasoning within GeoKGs.

The scalability constraints of GeoSPARQL’s on-the-fly spatial computations, and the separation of topological inferencing from broader semantic reasoning underscore the need for more scalable, semantically integrated approaches to spatial querying in GeoKGs.

3 Approach

To overcome the challenges that arise from relying on GeoSPARQL for spatial querying, topological predicates between spatial objects can be precomputed, which allows for more efficient direct lookup at query time. In the extreme case, this approach requires explicitly storing all topological relations between any combination of spatial objects, which quickly becomes infeasible for large or dynamic datasets. Instead, we seek a pragmatic compromise by precomputing only a much smaller set of topological relations, thus tailoring the *topological enrichment method* approach pioneered by Regalia et al. [48] and refined by KnowWhereGraph (KWG) [29, 56]. Just like KWG, we choose to leverage the S2 Geometry framework [50], which we elaborate on next, and explicitly represent it as part of the content of the GeoKG. Then, rather than precomputing topological relations between all kinds of geometric features, we only precompute them between the features and two types of common spatial reference entities (SREs) – S2 cells and administrative regions – to save space and increase retrieval efficiency. For that reason, we refer to this tailored approach by the name *SRE+Topology*. The precomputed relations are explicitly materialized in the graph to reduce the need for computationally expensive on-the-fly geometric computations during query execution.

S2 Geometry. Google’s S2 Geometry [50] defines a hierarchical and discrete global grid system that tessellates the Earth’s surface into a structured set of connected and well-aligned quadrilateral cells. These cells have geodesic edges and are organized into a nested hierarchy of cells with increasingly finer resolutions (levels). The hierarchy consists of 30 levels, where the average area ranges from $\sim 8.5 \cdot 10^7 km^2$ (level 0) to $\sim 0.74 cm^2$ (level 30). Each S2 cell is recursively subdivided into four cells at each subsequent level. S2 cells are sequentially ordered along a Hilbert space-filling curve, which projects the unit sphere’s surface onto six cube faces. Each cell is uniquely identified by a `S2CellID` that encodes its hierarchical level and its position on the Hilbert curve.

Semantic Representation of S2 Geometry. GeoSPARQL-compliant RDF databases such as GraphDB support S2 Geometry neither conceptually nor via specialized indexing data structures. To take advantage of S2 Geometry in a GeoKG, KWG represents S2 cells and their interrelations explicitly in the graph using a minimal ontology [54, 56] with a set of spatial relations that mirror those of GeoSPARQL as shown and described in Figure 1. The geometry of each `kwg-ont:S2Cell` is represented as a polygon with four vertices. To account for the hierarchical structure of S2 Geometry, `kwg-ont:S2Cell` is specialized into

PFAS are a group of thousands of synthetic chemicals associated with various health issues in humans. Known as “forever chemicals”, they are highly persistent in the environment because their strong carbon-fluorine bonds resist degradation, allowing them to accumulate in air, soil, and water. Exposure to PFAS is associated with various adverse health effects, including elevated cholesterol levels, reduced vaccine response in children, liver enzyme changes, pregnancy complications, and elevated risk of kidney and testicular cancer [1, 55]. PFAS contamination arises from various sources, such as chemical plants, landfills, wastewater, biosolids applied as agricultural fertilizers, airports, and firefighting training sites. Non-point sources, including spills and atmospheric deposition, further contribute to the widespread environmental dispersion of PFAS. This ubiquity, combined with its significant health and environmental risks, requires robust, integrative monitoring and mitigation efforts.

4.1 Use Cases: Environmental Contamination with PFAS

PFAS fate and transport in the environment involve complex processes, and testing is costly, resulting in many unanswered questions for experts and decision-makers working to identify, mitigate, and remediate contamination. To assist them, SAWGraph merges public PFAS-related datasets from federal and state agencies into a single GeoKG. This design is based on competency questions gathered from discussions with potential users, leading to three main use cases, each accompanied by example competency questions:

1. *Find Testing Results and Gaps*: Find PFAS test results from drinking water, groundwater, and agricultural soils and identify coverage gaps in testing. E.g.,
 - What water bodies are near potential contamination sources?
 - Where is PFAS contamination highly likely, but no testing has occurred?
2. *Contaminant Tracing*: Trace how PFAS may have been transported via spatial and hydrological connections from known or suspected contamination sources. E.g.,
 - What potential point sources are upstream from observed high PFAS concentrations in water, soil, or biota?
 - Do the test results downstream from a potential point source show measurable contamination in the surrounding environment?
3. *Assessing Risk and Identifying Vulnerable Populations*: Identify what areas and populations are likely to be impacted the most by PFAS contamination to support equitable access to testing capacities and mitigation resources. E.g.,
 - Which county subdivisions have high PFAS contamination and highly vulnerable populations based on economic and demographic indicators?
 - Which areas rely on private wells and have a high risk of groundwater contamination?

Answering these competency questions requires a range of spatial analysis operations, including proximity analysis, overlay analysis, and hydrographic network analysis. In Section 5, we will demonstrate the implementation and chaining of these operations within SPARQL queries using the Contaminant Tracing use case as an example. Prior to this, we will explain the construction of the graphs that comprise SAWGraph, including the datasets, ontologies, and precomputed topological links used in the process.

4.2 Datasets

The various use cases require ingesting and linking a diverse range of datasets, which are summarized in Table 1. In order to support modular reuse of the data and speed up queries that require only a small portion of the data, the data is divided into four thematically distinct

■ **Table 1** Examples of the thematic datasets integrated in SAWGraph.

Theme	Example Dataset	Description	Source
Contaminant Testing and Release Data	Safe Drinking Water Information System (SDWIS)	PFAS testing results for drinking water	EPA
	Environmental and Geographic Analysis Database (EGAD) [42]	state test results in surface and ground water and biota	Maine Dept. of Env. Protection
Facilities & Industries	Facility Registry Service	landfills, airports, defense sites, etc.	EPA
Hydrological Features	National Hydrography Dataset (NHD)	streams, surface water bodies, aquifers	USGS
	Water Well Database [40]	private water wells	Maine GS
Chemical Informatics	CompTox	chemical formula, structural identifiers, toxicity	EPA
Environm. and Social Context	Soil Survey	soil composition	USDA via KWG
	Census and American Community Survey	demographics	Census Bureau via Datacommons

knowledge graphs, which correspond to the first four data themes in Table 1: PFAS KG, FIO (Facilities and Industries) KG, Hydrology KG, and CompTox (Chemical Informatics) KG. They are supplemented by a fifth graph, the Spatial KG, which captures the S2 Geometry as well as administrative regions and serves as the spatial bridge across the graphs. Through federated querying – as illustrated in Section 5 – SAWGraph can access other GeoKGs, such as Geoconnext [13], KWG [29], and DataCommons [11], to retrieve additional environmental or social context information.

4.3 Ontologies

To structure the knowledge graphs, five connected and extensible OWL 2 ontologies were developed. They are shared at <https://github.com/SAWGraph> and form the semantic backbone of the five SAWGraph KGs: a contaminant ontology (ContaminOSO [21]; `coso:` for the PFAS KG), a facilities and industries ontology (`fio:`, FIO KG), an integrated hydrology ontology (multiple namespaces, Hydrology KG), a PFAS chemistry ontology (`comptox:`, CompTox KG), and the spatial ontology [57] summarized in Figure 1 (`kmg-ont:` and `spatial:`, Spatial KG). The namespaces utilized in the ontologies and SPARQL queries in Section 5 are listed in Table 2 in the Appendix, with their key upper-level classes and relations shown in Figure 2. These ontologies adopt and extend existing standardized ontologies as much as possible. COSO [21], for example, builds on the SOSA [59, 28], QUDT [24], and STAD [60] ontologies, while the hydrology ontology brings together multiple existing hydrology ontologies, including HY_Features [12], GWML2 [5, 22], and HyFO [6, 7, 18]. Both ContaminOSO and FIO have been newly developed specifically to support the SAWGraph project [19], but are made available for reuse by other Proto-OKN projects and other GeoKGs.

4.4 Implementation of the SRE+Topology Approach

SAWGraph extends KWG’s spatial ontology by introducing `spatial:connectedTo` as a property that subsumes all spatial contact relations (i.e., all topological relations except `kmg-ont:sfDisjoint`) and by adding meta-relations (e.g. declaring inverses) between them [57]. This additional semantic context is particularly useful for filtering data when

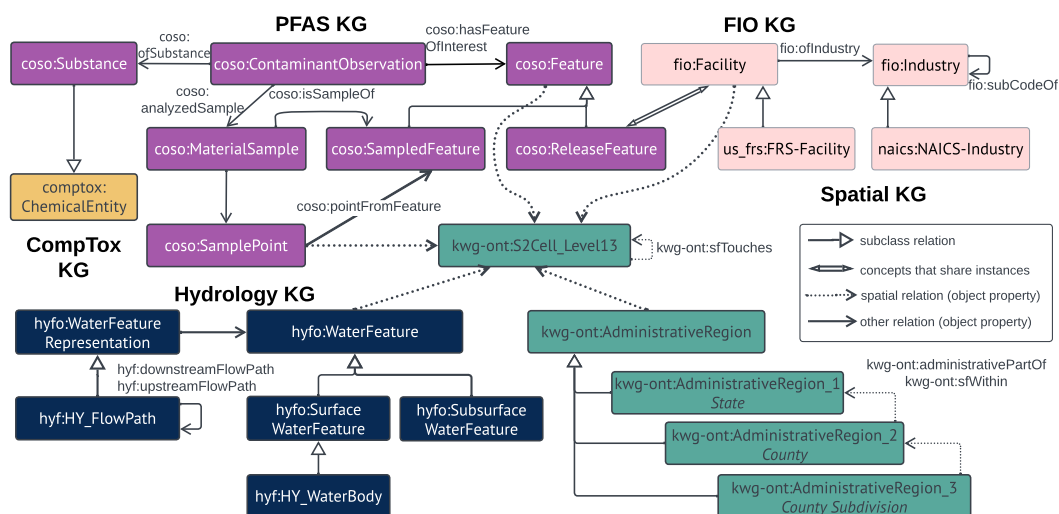


Figure 2 Conceptual overview of the five connected knowledge graphs that comprise SAWGraph and the ontologies they use. Each color represents one KG and its underlying core ontology, for which key high-level classes and relations are shown.

more precise topological relationships are not required. For instance, a water body may be represented as a point feature within a county or a polygon feature overlapping the county; both scenarios can be generalized as the water body being spatially connected to the county.

A key challenge in utilizing the SRE+Topology approach is managing the trade-off between storage and query efficiency. For example, materializing the topological relations between features and S2 cells across multiple levels of resolution is not feasible because the number of stored triples grows quadratically with the number of features (including S2 cells). To address this, we only precompute topological relations with two sets of static entities – level 13 S2 cells and level 3 administrative regions (i.e. county subdivisions in the US) – so that each point from a feature’s vector representation produces at most two triples that instantiate topological relations.

From S2 Geometry, SAWGraph only utilizes S2 cells of level 13. They span $\sim 0.76\text{--}1.59\text{ km}^2$ with an average area of 1.3 km^2 in the continental United States. This resolution strikes a balance between spatial granularity and computational and storage efficiency. It is well-suited for regional-scale analyses, particularly for monitoring environmental phenomena. KWG already included level 0–2 administrative regions (countries, states and counties) from the GADM dataset [15] and their precomputed topological relations with S2 cells. SAWGraph adds the level 3 administrative regions with the relation `kwg-ont:administrativePartOf` capturing how they are nested inside coarser administrative regions, which supports efficient lookups of geospatial features by any administrative regions up to level 3. For SAWGraph, the level 3 administrative regions and level 13 S2 cells are the only spatial reference entities for which topological relations with all other features are precomputed and materialized.

5 The Contaminant Tracing Case Study

PFAS contamination pathways are complex, often involving significant movement through water, air, and soil, and accumulating in unexpected locations. Better understanding how PFAS enters and moves through environmental systems is crucial for identifying exposure

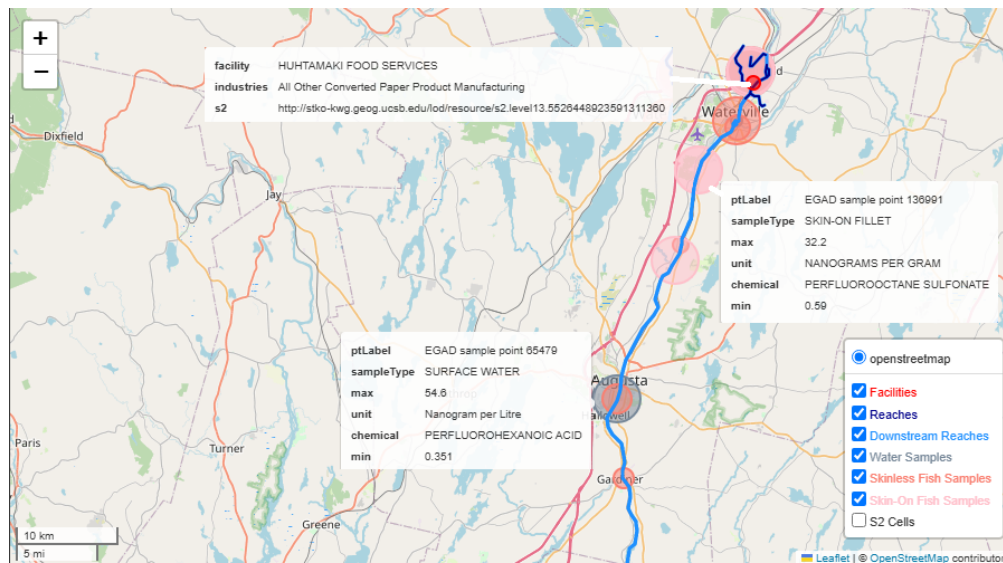


Figure 3 Interactive visualization of fish tissue and surface water sampling results downstream of paper manufacturing facilities in Maine, zoomed in on one facility close to the Kennebec River. The radii of the sampling results correspond to the highest concentration across all PFAS detected at the point. The full map of Maine is provided as Figure 6 in the Appendix.

and developing targeted interventions. A key way PFAS spreads is through hydrological systems, such as rivers and groundwater [51, 33]. Contaminated water can infiltrate drinking water supplies, agricultural irrigation, and aquatic ecosystems, creating multiple exposure risks for humans, livestock, and wildlife. These pathways complicate source attribution, which is essential for effective mitigation and remediation and for the design of targeted regulations, such as restrictions on PFAS use in specific industries. In addition, improving our understanding of contamination pathways aids in developing accurate fate and transport models that simulate the movement of PFAS in environmental systems.

Many federal or state agencies are charged with monitoring contaminants like PFAS in water, food and the environment. To fulfill this mission, they regularly analyze water, soil and tissue samples for contamination. For example, Maine DEP and DACF have analyzed hundreds of groundwater and surface water samples but also samples of fish, seafood, other animal, and soil for PFAS. The collected data were used for prototyping SAWGraph. For the purpose of this paper, we will demonstrate the utility of SAWGraph and its implementation of the SRE+Topology approach to gain insights into source-to-impact pathways and the role that particular industries or facilities play in PFAS contamination, focusing on two particular analytic questions that evaluate the role of converted paper product manufacturing facilities – some of which might have used PFAS for coated paper products or for the smooth operation of their machinery – as PFAS point sources: *What does the data show about fish tissue and surface water contamination downstream of converted paper product manufacturing facilities in Maine?* (Question 1) Figure 3 shows the resulting map. We also explore a follow-up question: *Which areas downstream of paper manufacturing facilities are not in a public water service area?* (Question 2)

Answering these questions requires accessing multiple graphs to link industrial facilities to PFAS observations through the hydrological network and spatial graph, as illustrated by the connections between the graph's key concepts in Figure 2. Each question can be expressed as

■ **Query Segment 1** Use of spatial intersection (Blocks B1a, B1b), spatial proximity (B1b, B1c), and network tracing query (B1c) to locate facilities by industry (converted paper product manufacturing) and administrative region (Maine); to retrieve their S2 cell neighborhoods (S2 cell and all eight neighbors) and the stream reaches flowing through those neighborhoods; and to find all downstream stream reaches and their S2 cells.

```

1 SELECT * WHERE {
2   SERVICE <repository:FIO> { # B1a: Retrieve facilities and their locations
3     ?industry fio:subcodeOf naics:NAICS-3222 . # Converted Paper Product
4       Manufacturing
5     ?facility_iri a fio:Facility ; # IRI (unique identifier) of each facility
6     rdfs:label ?facility_label ; # Human-readable label (name) of each facility
7     fio:ofIndustry ?industry ; # Filter to selected industry
8     geo:hasGeometry/geo:asWKT ?facility_wkt ; # Facility geometry as WKT string
9     spatial:connectedTo ?s2_cell ; # S2 cell that the facility is located in
10    spatial:connectedTo ?countysub . # County subdivision the facility is in
11  }
12  SERVICE <repository:Spatial> { # B1b: spatially filter to State of Maine (USA.23)
13    ?countysub a kwg-ont:AdministrativeRegion_3 ;
14    kwg-ont:administrativePartOf+ kwgr:administrativeRegion.USA.23 .
15    ?s2_cell a kwg-ont:S2Cell_Level13 .
16    ?s2_neighborhood kwg-ont:sfTouches | owl:sameAs ?s2_cell ; # Facility S2 cell
17    neighborhood (S2 cell and its 8 neighbors)
18    geo:hasGeometry/geo:asWKT ?s2_wkt . # S2 cell geometries for visualization
19    ?s2_ds_reach a kwg-ont:S2Cell_Level13 ; # Downstream S2 cells
20    geo:hasGeometry/geo:asWKT ?s2_ds_reach_wkt . # Downstream S2 geometries
21  }
22  SERVICE <repository:Hydrology> { # B1c: tracing hydrological network downstream
23    ?reach a hyf:HY_FlowPath ;
24    spatial:connectedTo ?s2_neighborhood ; # Stream reaches crossing the
25    facility S2 neighborhoods
26    hyf:downstreamFlowPath+ ?ds_reach . # Downstream stream reaches
27    ?ds_reach geo:hasGeometry/geo:asWKT ?ds_reach_wkt ; # Stream reach geometries
28    spatial:connectedTo ?s2_ds_reach . # S2 cells for downstream stream reaches
29  }
30  ...

```

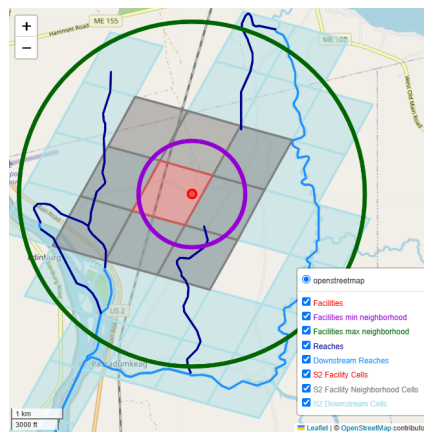
a single SPARQL query but for validation and visualization purposes we often divide them. In this paper, the example query is divided into segments that exemplify important classes of geospatial operations familiar to GIS users. Altogether, we use five basic operations that are essential for constructing a wide range of complex geospatial workflows, namely:

1. **Spatial intersection/filtering:** Find contamination point sources (e.g., converted paper product manufacturing facilities) that are within the target region (e.g. Maine).
2. **Proximity:** Find all stream reaches that are near these facilities (e.g., within 1-2km²).
3. **Network tracing and distance:** Trace all stream reaches downstream.
4. **Proximity and spatial intersection:** Find all PFAS observations from surface water and fish tissue samples near any of the downstream stream reaches.
5. **Vector overlay:** Find contaminated areas that are outside public water service areas.

We describe the logic and SPARQL implementation of these operations next.

5.1 Spatial Intersection: Find facilities in the area of interest

The first query retrieves all industrial facilities classified as converted paper product manufacturing industries (Block B1a of Query Segment 1) using the FIO graph and then spatially filtering them to those located in the state of Maine (B1b) using the Spatial graph. More specifically, B1a first retrieves all subindustry codes from the broad group of the NAICS



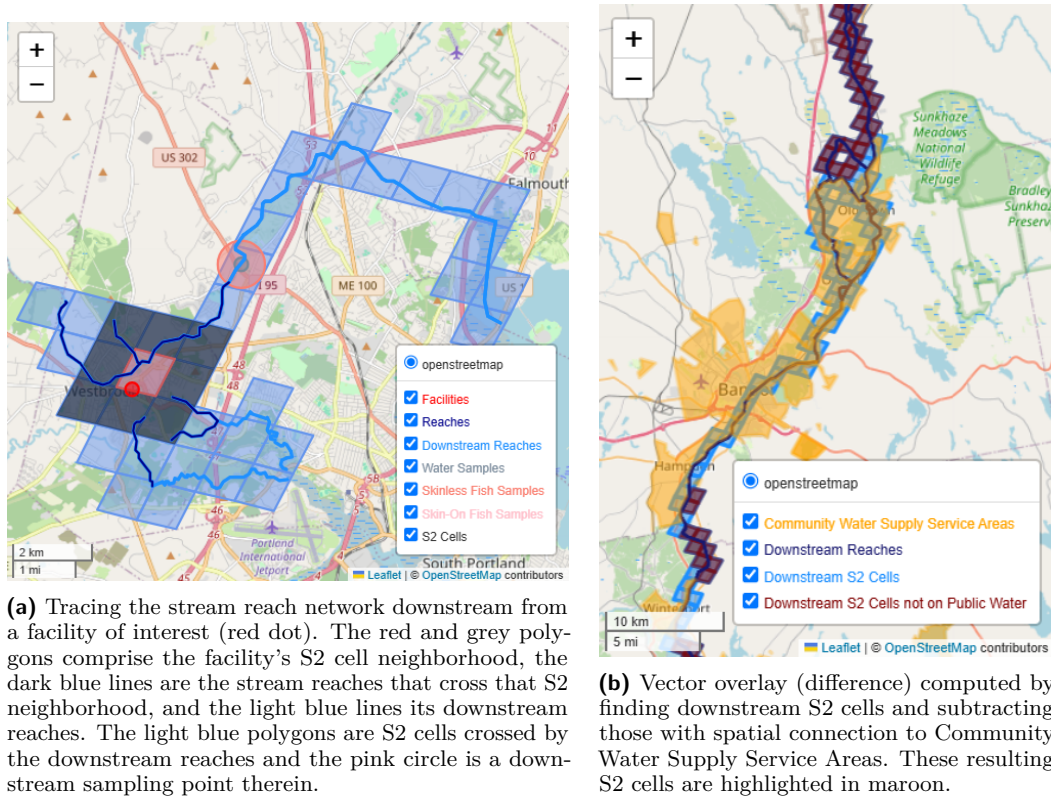
■ **Figure 4** Example facility with its S2 cell (red) and S2 neighborhood (grey). For comparison, the purple and green circles show what would be included in a standard proximity search with radius 1015 m or 3264 m, which correspond to the minimum length of an S2 cell's side or twice its longest diagonal. See the main text for more details.

industry code 3222 (i.e., converted paper product manufacturing) because facilities are typically associated with the most fine-grained industry labels available. These are then used to identify facilities whose industry code matches any of those subindustries (lines 4-6). The facilities are retrieved along with their geometries (line 7) and the precomputed S2 cells and county subdivisions (`AdministrativeRegion_3`) they are in (lines 8, 9). Block B1b leverages the hierarchical structure of the administrative regions from the Spatial KG to identify which county subdivisions are within the state of Maine (identified by its URI `kwgr:administrativeRegion.USA.23`, lines 12, 13) to eliminate facilities outside of Maine. The precomputed topological relations between S2 cells (`connectedTo` and `sfTouches`) suffice for the spatial filtering needs here, thereby ensuring quick query responses.

5.2 Spatial Proximity: Find nearby stream reaches

Tracing where contaminants emitted by the facilities may be transported via surface water flow requires first locating which stream reaches (i.e., hydrological flow segments, which are represented as `hyf:HY_FlowPath` using the `HY_Features` ontology [12]) are in proximity to the identified paper manufacturing facilities. If we were to only consider stream reaches that intersect the S2 cell where a facility is located, nearby reaches could be missed when the facility is close to the border of its encompassing S2 cell. To perform proximity or similar buffering operations, it is better to leverage the metric implicitly built into the S2 grid, which is defined by the fairly uniform sizes of level 13 cells (or cells of any particular level). For example, we can approximate the neighborhood of facilities by including the eight neighboring cells of the S2 cell where a facility is located. If a larger distance is desired, one could expand that to the additional 16 neighbors of the neighbors, and so on.

By including the eight S2 neighbors, we guarantee to find all stream reaches within a radius equal to the length of the shortest side of the S2 cell, as illustrated in Figure 4 using the shortest side of the center (red) S2 cell as radius. A circle of this radius, centered at any point in the center S2 cell, will always be entirely within the S2 neighborhood. The green circle has a radius equal to twice the longer diagonal of the center S2 cell to guarantee that the entire eight-cell S2 neighborhood is fully included no matter where the circle is centered



■ **Figure 5** Example map results illustrating the network tracing and vector overlap operations.

within the center S2 cell. Stream reaches outside it will never be deemed “near” the facility by the S2-based approach. Thus, the radii of the red and green circles describe the lower and upper bound of the proximity operation’s spatial precision.

Because our approach is agnostic of where a feature is within an S2 cell, it cannot search within a fixed radius around a point location but approximates the search area using grid cells. It limits spatial precision but gains efficiency because it avoids the need to compute distances or buffers on-the-fly. At query time, the set of S2 cells describing the proximal area can be retrieved from the Spatial KG and passed on to the Hydrology KG for retrieving the stream reaches that intersect those S2 cells (Query Segment 1, B1c, line 21).

5.3 Network Tracing and Network Distance: Trace stream reaches

The identified stream reaches from Query Segment 1 (denoted by variable `?s2_ds_reach` and shown as dark blue lines in Figure 5a) serve as starting points for our network tracing task. The stream reaches are the smallest hydrological flow segments connected to one another via the relation `hyf:downstreamFlowPath` and its inverse `hyf:upstreamFlowPath` in SAWGraph, which are based on NHD’s downstream and upstream relations to define a flow direction. They allow the construction of longer flow paths, which are directed paths that each consist of a sequence of one or more stream reaches and can be traced upstream (i.e., from a sink to a source) or downstream (i.e., from a source to a sink). For our question, Block B1c of Query Segment 1 uses the Hydrology graph to trace the stream reaches downstream (light blue lines in Figure 5a) by exploiting the transitive closure of the `hyf:downstreamFlowPath` relation using SPARQL’s transitive path operator “+” (line 23).

The same effect would be achieved by defining `hyf:downstreamFlowPathTC` as a transitive superproperty thereof in the ontology (see [20]), which is propagated and prematerialized during graph construction and, thus, even faster. Either approach provides a structured way to navigate the hydrological network and simulate flow paths originating from a given starting point.

It may not always be desirable to consider *all* stream reaches downstream of a given feature. Because the KG stores the length of each reach, it is possible to limit downstream reach to those within a chosen maximum flow path length. This can be accomplished by adding the subquery shown in Query Segment 2 to Block B1c of Query Segment 1 along with a filter to set the maximum length. The subquery takes a reach (`?reach`) that is near a facility along with any of its downstream reaches (`?ds_reach`), and then sums the lengths of all intermediate stream reaches (`?f1`). Because each stream reach is defined as downstream of itself (for this specific purpose), the total distance includes the entire lengths of both ends of the flow path. In the example, only flow paths shorter than 20 km are returned.

These kinds of tracing analyses can be expanded, for example, by using the S2 cells retrieved in Query Segment 1 to also identify potential hydrological connectivity – or at least proximity – between contaminated surface water bodies and groundwater aquifers. This could further improve contaminant tracing by locating groundwater resources that may be infiltrated by PFAS from nearby contaminated stream reaches.

5.4 Proximity and Spatial Intersection: Find relevant PFAS results

The final step in answering Question 1 focuses on retrieving PFAS-related data, such as water quality measurements or fish tissue contamination levels, from samples collected along the downstream reaches of the hydrological network. Since sampling observations and hydrological datasets are in distinct thematic layers, we can only establish meaningful correlations by first spatially linking them via the S2 cells as spatial reference entities. However, stream reaches are often represented as 1-dimensional geometric approximations of a water body's central flow path, which exclude the width and area of the river. Consequently, sampling points, represented as 0-dimensional geometries, that were originally within the river's boundaries may no longer intersect with the simplified line geometries. One approach to mitigate this issue is to apply a buffer around the stream reaches, approximating the river's extent and improving the accuracy of the intersection. However, it may still miss

■ **Query Segment 2** An optional subquery for Block B1c from Query Segment 1 to limit downstream navigation to a specific distance (20km in this example).

```

1  ... { SELECT ?reach ?ds_reach (SUM(?f1_length) AS ?path_length) WHERE {
2      ?reach a hyf:HY_FlowPath ;
3      spatial:connectedTo ?s2_neighborhood ; # Stream reaches crossing
4      the facility S2 cells
5      hyf:downstreamFlowPath+ ?f1 . # Stream reaches between those
6      crossing a facility S2 cell and some downstream reach
7      ?f1 a hyf:HY_FlowPath ;
8      hyf:downstreamFlowPath+ ?ds_reach ; # Last stream reach in a chain
9      starting from a stream reach crossing the facility S2 cells
10     nhdplusv2:hasFlowPathLength/qudt:quantityValue/qudt:numericValue
        ?f1_length . # Flow path length
    } GROUP BY ?reach ?ds_reach
    } FILTER (?path_length < "20.0"^^xsd:float)
    } ...

```

sampling points located just outside along the shore. Another approach is to calculate the distance from each sampling point to the nearest stream reach and retrieve points within a reasonable threshold. However, both methods involve computationally expensive geometric operations, which can be impractical whenever the datasets become larger.

To overcome these limitations, our solution (see Query Segment 3) again leverages the S2 cells (variable `?s2_ds_reach` from Query Segment 1) that intersect the downstream reach segments. These S2 cells act as approximate spatial buffers, enabling efficient filtering of PFAS sampling data without the need for computationally intensive geometric calculations. The query retrieves all sampling observations whose sampling points are within those S2 cells (lines 3–4). Lines 6 and 7 then retrieve information about their material sample type (e.g., water or fish) and Block B3b accesses the contamination observation results using the SOSA observation-measurement-result pattern [59].

5.5 Vector Overlay: Find impacted areas without public water supply

In addition to supporting spatial filtering and proximity tasks, the SRE+Topology approach also supports simplified and efficient proxies for more expensive spatial overlay operations such as polygon intersection, union and difference. We demonstrate this functionality by determining which of the reaches downstream from potentially polluting facilities are inside (*intersection*) or outside (*difference operation*) of community water supply service areas to address Question 2 introduced at the beginning of Section 5. It helps prioritize PFAS testing in areas without public drinking water access where residents typically rely on private wells that may be affected by the contaminated water table. Analogous to Query Segment 3, we take the S2 cell neighborhoods of all downstream reaches (`?s2_ds_reach`) as an approximate buffer, and overlay them with the (precomputed) S2 cells that overlap with any community water supply service area to determine the difference between the two sets of S2 cells to avoid computationally expensive spatial calculations.

This analysis is just one of many; Query Segment 3 could be expanded further by adding other environmental variables, such as soil type, precipitation, and land use, via federated querying of external graphs to put the contamination results (encoded by the variable `?measure`) in context. It could guide testing and monitoring strategies by examining the

■ **Query Segment 3** Finding PFAS sampling observations in the proximity of the stream reaches downstream from the paper manufacturing facilities identified in Query Segment 1 by using the S2 cell neighborhoods around the reaches.

```

1 ... #Continued from Query Segment 1 and 2
2 SERVICE <repository:PFAS> { # B3a: Find sampling points in surface water
3   ?sample_point a coso:SamplePoint ; # Find sampling points within ...
4   spatial:connectedTo ?s2_ds_reach ; # ... downstream S2 cells
5   geo:hasGeometry/geo:asWKT ?sample_point_wkt . # Get sampling point geometry
6   ?material_sample coso:fromSamplePoint ?sample_point ;
7   coso:ofSampleMaterialType ?sample_type . # Identify type of sample
8   # B3b: Identify analyzed PFAS substance and measurement value
9   ?observation coso:analyzedSample ?material_sample ; # Get each observation
10  coso:ofSubstance ?substance ; # Get PFAS chemical analyzed
11  coso:hasResult ?measure . # Get result of the observation
12  ?measure qudt:quantityValue ?quantity_v . # Get quantity from result
13  ?quantity_v qudt:numericValue ?value ; # Numeric value of the quantity
14  qudt:unit ?unit. # Unit of the quantity
15 }
16 }
```


■ **Query Segment 4** Spatial overlay for finding downstream reaches outside public drinking water service areas.

```

1 ... #Continued from Query Segment 1
2 SERVICE <repository:Hydrology> { #B4b: Subtract public drinking water areas
3   MINUS { ?s2_ds_reach spatial:connectedTo ?pws .
4           ?pws a us_sdwis:PWS-ServiceArea .} }

```

correlations highly contaminated stream reaches exhibit with respect to, e.g., agricultural activity, population density, or industrial land use; or prioritize interventions by ranking regions by vulnerability based on observed contamination, environmental factors, and human exposure risks.

5.6 Comparison to GeoSPARQL Operations

For comparison we also implemented and executed Question 1 using on-the-fly GeoSPARQL functions and predicates to perform the same analysis though obtaining the precise rather than spatially approximated results². The geometries of our features are stored in 3-D coordinates (latitude longitude WGS84), and therefore we use a proximity distance of 0.014 arc degrees, which is equivalent to approximately 1119.06 m in our study area at 44 degrees North latitude. To perform the equivalents of Query Segments 1 and 3 in GeoSPARQL we use a distance search (`geof:distance`) on facilities within Maine (`geo:sfWithin`) to find nearby stream reaches, follow them downstream, and then buffer downstream reaches to find sampling points within the downstream reach buffer. This query completes in 165s, compared to our equivalent S2-based query in Section 5.4, which completes in 21s when executed on the same server under the same conditions. The question as defined is limited to only converted paper manufacturing facilities in Maine, which encompasses only 10 facilities. When we expand this search to all facilities in Maine in industries suspected of using PFAS, which encompasses a total of 354 facilities, the GeoSPARQL query completes in approximately 84 minutes (1h 23m 55s) while the equivalent S2-based query takes less than 11 minutes (10m 39s). Both S2-based queries achieve an eightfold – almost an order of magnitude – speedup. More importantly, these improvements do not rely on using any internal quadtree or other specialized indexing data structure for encoding the S2 geometry. Thus, we would expect comparable performance of the SRE+Topology approach in other RDF graph databases regardless of whether they provide any kind of geospatial indexing or GeoSPARQL support. A much more comprehensive comparison will be part of future work.

6 Summary and Discussion

We have demonstrated how the SRE+Topology approach supports efficient execution of advanced geospatial questions, such as about environmental contamination, directly in a GeoKG without the need for specialized reasoners, spatial indexing, or the GeoSPARQL geometric operations. Our example questions about environmental contamination combine network analysis with intersection, proximity, and overlay operations. For example, knowledge about the hydrological network for contaminant transport is leveraged together with proximity information and spatial intersections to identify downstream contamination risks.

² The original and optimized queries and their GeoSPARQL equivalents are available from <https://github.com/SAWGraph/public/tree/main/UseCases/UC3-Tracing/UC3-CQ15/GIScience2025-queries>.

Executing such advanced geospatial analysis questions in a GeoKG using GeoSPARQL operations instead of the precomputed topological relations would require spatial indexing and/or expensive spatial computations for geometric overlays, buffering, and topological analysis across features from multiple geospatial data layers. In the SRE+Topology approach, these spatial tasks are addressed in a unified way that relies entirely on precomputed topological links between different features and S2 cells, eliminating the need for resource-intensive geometric operations at query time. Querying a large GeoKG via these links maintains computational efficiency while enabling complex analyses across large datasets and extensive geographic ranges. The SRE+Topology approach facilitates the construction of these queries within and across graphs using standard SPARQL constructs only, that is, without the need for GeoSPARQL, thereby democratizing geospatial analysis via GeoKGs. Moreover, the proposed approach integrates the semantic representation afforded by GeoKGs with the analytic capabilities afforded by conventional GIS.

Furthermore, the SRE+Topology approach allows sharing spatial reference entities (SREs), such as S2 cells and administrative regions, across separate graphs. It offers a robust mechanism to distribute data into separate thematic GeoKGs while ensuring their spatial compatibility. Thereby, some of the scalability challenges related to graph construction, maintenance, storage, and querying experienced in KnowWhereGraph – which was constructed as a single monolithic GeoKG – can be overcome. With SRE+Topology, different thematic information, such as hydrological, environmental, or socio-economic information, can be stored in separate GeoKGs, each maintained by their respective data producers or owners. Through the precomputed topological relations between features from these independent graphs and the shared SREs, the GeoKGs can be queried jointly using SPARQL’s federation construct (`SERVICE`). This modular and distributed architecture supports the growth of these graphs and helps accommodate more diverse and dynamic spatial datasets.

The SRE+Topology approach is, by design, a compromise between a full explicit representation of topological relationships, which would be impractical, and the classical approach of computing spatial queries on-the-fly. This design naturally comes with some drawbacks, which we can only outline here but that require future study. The first is the computational and storage overhead caused by precomputing and storing the intersections of all features in the thematic GeoKGs with the spatial reference entities. The number of additional triples for representing the SREs is constant, thus it is critical to carefully select suitable reference entities. In SAWGraph, we choose level 13 S2 cells and level 3 administrative regions to strike a balance between spatial granularity and computational demands (storage and query processing times). The number of triples for representing the topological relations only grows linearly in terms of the number of geographic features stored across all thematic layers and can be distributed as well. Efficient precomputation may also be more problematic for highly dynamic datasets, as any updates require recomputing the stored topological relations, adding potentially significant maintenance overhead.

A related limitation concerns the afforded spatial granularity and thus spatial precision, in particular for queries that require precise geometric measures, such as distances or buffers. The supported spatial granularity is directly tied to the choice of S2 cell or administrative region level used as SREs. Rather than switching everything to finer-grained S2 cells (or other SREs), which would rapidly increase storage needs, a more flexible approach could leverage the hierarchical relations (e.g., `kwg-ont:sfContains`) between different levels of SREs to allow topologically linking thematic features to the level that best reflects the granularity of a specific thematic dataset. Another option is a hybrid approach, where the SRE+Topology approach is used to narrow the set of potential features of interest to a small subset of all

features (e.g., all PFAS sample locations within the S2 neighbors that overlap a stream reach) before applying precise geometric operations, such as a distance function, to calculate the exact distance of each such sample location from the stream reach to determine whether to include or exclude the location. However, suitable querying approaches require careful design and testing to verify that they actually are more storage and/or time efficient. Finally, some spatial operations, such as those that construct new polygons from the intersection of existing polygons rather than just determining whether they intersect, cannot be easily implemented using only the SRE+Topology approach but would require a hybrid approach.

References

- 1 Agency for Toxic Substances and Disease Registry (ATSDR). Per- and polyfluoroalkyl substances (PFAS) and your health, November 2024. URL: <https://www.atsdr.cdc.gov/pfas/about/health-effects.html>.
- 2 Chaitan Baru, Martin Halbert, Lara Campbell, Tess DeBlanc-Knowles, Jemin George, Wo Chang, Adam Pah, Douglas Maughan, Ilya Zaslavsky, Amanda Stathopoulos, et al. Open knowledge network roadmap: Powering the next data revolution. Technical report, National Science Foundation, 2022.
- 3 Robert Battle and Dave Kolas. GeoSPARQL: Enabling a geospatial semantic web. *Semantic Web Journal*, 3(4):355–370, 2011.
- 4 Dan Brickley and Ramanathan Guha. RDF vocabulary description language 1.0: RDF schema. W3C recommendation, W3C, February 2004. URL: <https://www.w3.org/TR/2004/REC-rdf-schema-20040210/>.
- 5 B. Brodaric, E. Boisvert, Chery. L., P. Dahlhaus, S. Grellet, A. Kmoch, F. Letourneau, J. Lucido, B. Simons, and B. Wagner. Enabling global exchange of groundwater data: GroundWaterML2 (GWML2). *Hydrogeology Journal*, 26(3):733–741, 2018. doi:10.1007/s10040-018-1747-9.
- 6 Boyan Brodaric and Torsten Hahmann. Towards a foundational hydro ontology for water data interoperability. In *11th International Conference on Hydroinformatics (HIC-2014)*, pages 2911–2915, 2014.
- 7 Boyan Brodaric, Torsten Hahmann, and Michael Gruninger. Water features and their parts. *Applied Ontology*, 14(1):1–42, 2019. doi:10.3233/A0-190205.
- 8 Ying Chen. *Enhancing Spatial Query Efficiency Through Dead Space Indexing in Minimum Bounding Boxes*. PhD thesis, University of Waterloo, 2024.
- 9 James Cheng, Yiping Ke, and Wilfred Ng. Efficient query processing on graph databases. *ACM Transactions on Database Systems (TODS)*, 34(1):1–48, 2009. doi:10.1145/1508857.1508859.
- 10 Eliseo Clementini, Paolino Di Felice, and Peter van Oosterom. A small set of formal topological relationships suitable for end-user interaction. In David Abel and Beng Chin Ooi, editors, *Advances in Spatial Databases*, pages 277–295. Springer, 1993. doi:10.1007/3-540-56869-7_16.
- 11 Data Commons. Data commons 2025. URL: <https://datacommons.org>.
- 12 Irina Dornblut and Robert Atkinson. HY_Features: a geographic information model for the hydrology domain. Technical Report GRDC 43r1, Global Runoff Data Centre, November 2013.
- 13 Martin Doyle and Kyle Onda. Internet of water: Research and development toward a linked data system and foundational knowledge network for the internet of water. Technical report, NC WRI, 2023.
- 14 Sebastián Ferrada, Benjamin Bustos, and Aidan Hogan. Similarity joins and clustering for SPARQL. *Semantic Web*, 15(5):1701–1732, 2024. doi:10.3233/SW-243540.
- 15 GADM maps and data. URL: <https://gadm.org/index.html>.

- 16 George Garbis, Kostis Kyzirakos, and Manolis Koubarakis. Geographica: A benchmark for geospatial RDF stores. In *12th International Semantic Web Conference (ISWC 2013)*, pages 343–359. Springer, 2013.
- 17 Torsten Hahmann. Ontology. In B.S. Daya Sagar, Qiuming Cheng, Jennifer McKinley, and Frits Agterberg, editors, *Encyclopedia of Mathematical Geosciences*, pages 1–5. Springer, 2021. doi:10.1007/978-3-030-26050-7_231-1.
- 18 Torsten Hahmann and Boyan Brodaric. The void in hydro ontology. In *12th International Conference on Formal Ontology in Information Systems (FOIS-12)*, pages 45–58. IOS Press, 2012. doi:10.3233/978-1-61499-084-0-45.
- 19 Torsten Hahmann, Pascal Hitzler, Hande Küçük McGinty, Ganga Hettiarachchi, Onur Apul, et al. Safe Agricultural Products and Water Graph (SAWGraph): An Open Knowledge Network to Monitor and Trace PFAS and Other Contaminants in the Nation’s Food and Water Systems. URL: <https://sawgraph.github.io/>.
- 20 Torsten Hahmann and David K Kedrowski. An ontology and geospatial knowledge graph for reasoning about cascading failures. In *16th International Conference on Spatial Information Theory (COSIT 2024)*, pages 21:1–9. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2024. doi:10.4230/LIPIcs.COSIT.2024.21.
- 21 Torsten Hahmann, Katrina Schweikert, Shirly Stephen, and David Kedrowski. ContaminOSO: Ontological foundations and key design choices for an ontology for environmental contaminant data. In *25th International Conference on Formal Ontology in Inf. Systems (FOIS-25)*. IOS Press, 2025 (to appear).
- 22 Torsten Hahmann and Shirly Stephen. Using a hydro-reference ontology to provide improved computer-interpretable semantics for the groundwater markup language (GWML2). *International Journal of Geographic Information Science*, 32(6):1138–1171, 2018. doi:10.1080/13658816.2018.1443751.
- 23 Pascal Hitzler, Bijan Parsia, Peter Patel-Schneider, and Sebastian Rudolph. OWL 2 Web Ontology Language Primer (Second Edition). <https://www.w3.org/TR/owl2-primer/>, 2012. <https://www.w3.org/TR/owl2-primer/>.
- 24 Ralph Hodgson, Paul J. Keller, Jack Hodges, and Jack Spivak. QUDT: Quantities, units, dimensions and types, 2012. URL: <https://qudt.org/>.
- 25 Weiming Huang, Syed Amir Raza, Oleg Mirzov, and Lars Harrie. Assessment and benchmarking of spatially enabled RDF stores for the next generation of spatial data infrastructure. *ISPRS International Journal of Geo-Information*, 8(7):310, 2019. doi:10.3390/IJGI8070310.
- 26 Theofilos Ioannidis. Geospatial RDF stores. In *Geospatial Data Science: A Hands-on Approach for Building Geospatial Applications Using Linked Data Technologies*, pages 221–240. Association for Computing Machinery, 2023.
- 27 Theofilos Ioannidis, George Garbis, Kostis Kyzirakos, Konstantina Bereta, and Manolis Koubarakis. Evaluating geospatial RDF stores using the benchmark Geographica 2. *Journal on Data Semantics*, 10(3):189–228, 2021. doi:10.1007/S13740-021-00118-X.
- 28 Krzysztof Janowicz, Armin Haller, Simon J.D. Cox, Danh Le Phuoc, and Maxime Lefrançois. Sosa: A lightweight ontology for sensors, observations, samples, and actuators. *Journal of Web Semantics*, 56:1–10, 2019. doi:10.1016/j.websem.2018.06.003.
- 29 Krzysztof Janowicz, Pascal Hitzler, Wenwen Li, Dean Rehberger, Mark Schildhauer, Rui Zhu, Cogan Shimizu, Colby Fisher, Ling Cai, Gengchen Mai, et al. Know, Know Where, KnowWhereGraph: A densely connected, cross-domain knowledge graph and geo-enrichment service stack for applications in environmental intelligence. *AI Magazine*, 43(1):30–39, 2022. doi:10.1609/AIMAG.V43I1.19120.
- 30 Peiquan Jin, Xike Xie, Na Wang, and Lihua Yue. Optimizing R-tree for flash memory. *Expert Systems with Applications*, 42(10):4676–4686, 2015. doi:10.1016/J.ESWA.2015.01.011.
- 31 J Michael Johnson, Tom Narock, Justin Singh-Mohudpur, Doug Fils, Keith Clarke, Siddharth Saksena, Adam Shepherd, Sankar Arumugam, and Lilit Yeghiazarian. Knowledge graphs to support real-time flood impact evaluation. *AI Magazine*, 43(1):40–45, 2022. URL: <https://ojs.aaai.org/index.php/aimagazine/article/view/19121>.

- 32 Milos Jovanovik, Timo Homburg, and Mirko Spasić. A GeoSPARQL compliance benchmark. *IS-PRS International Journal of Geo-Information*, 10(7):487, 2021. doi:10.3390/IJGI10070487.
- 33 Sudarshan Kurwadkar, Jason Dane, Sushil R Kanel, Mallikarjuna N Nadagouda, Ryan W Cawdrey, Balram Ambade, Garrett C Struckhoff, and Richard Wilkin. Per-and polyfluoroalkyl substances in water and wastewater: A critical review of their global occurrence and distribution. *Science of The Total Environment*, 809:151003, 2022.
- 34 Wenwen Li, Sizhe Wang, Sheng Wu, Zhining Gu, and Yuanyuan Tian. Performance benchmark on semantic web repositories for spatially explicit knowledge graph applications. *Computers, Environment and Urban Systems*, 98:101884, 2022. doi:10.1016/J.COMPENVURBSYS.2022.101884.
- 35 Jiajun Liu, Haoran Li, Yong Gao, Hao Yu, and Dan Jiang. A geohash-based index for spatial data management in distributed memory. In *22nd International Conference on Geoinformatics*, pages 1–4, 2014. doi:10.1109/GEOINFORMATICS.2014.6950819.
- 36 Junnan Liu, Haiyan Liu, Xiaohui Chen, Xuan Guo, Qingbo Zhao, Jia Li, Lei Kang, and Jianxiang Liu. A heterogeneous geospatial data retrieval method using knowledge graph. *Sustainability*, 13(4):2005, 2021.
- 37 Gengchen Mai, Yingjie Hu, Song Gao, Ling Cai, Bruno Martins, Johannes Scholz, Jing Gao, and Krzysztof Janowicz. Symbolic and subsymbolic GeoAI: Geospatial knowledge graphs and spatially explicit machine learning. *Transactions in GIS*, 26(8):3118–3124, 2022. doi:10.1111/TGIS.13012.
- 38 Gengchen Mai, Krzysztof Janowicz, Bo Yan, and Simon Scheider. Deeply integrating linked data with geographic information systems. *Transactions in GIS*, 23(3):579–600, 2019. doi:10.1111/TGIS.12538.
- 39 Gengchen Mai, Krzysztof Janowicz, Rui Zhu, Ling Cai, and Ni Lao. Geographic question answering: challenges, uniqueness, classification, and future directions. In *24th AGILE Conference on Geographic Information Science*, pages 8:1–21. Copernicus Publications, 2021. doi:10.5194/agile-giss-2-8-2021.
- 40 Maine Geological Survey (MGS). Water well database. <https://www.maine.gov/dacf/mgs/pubs/digital/well.htm>. Accessed 3 October 2023.
- 41 National Science Foundation. The Proto-OKN initiative, 2023. URL: <https://www.proto-okn.net/>.
- 42 Maine Department of Environmental Protection. Environmental and geographic analysis database EGAD. <https://www.maine.gov/dep/maps-data/egad/>. March 2024 release.
- 43 Open Geospatial Consortium. OGC GeoSPARQL - A Geographic Query Language for RDF Data. Open Geospatial Consortium, URL <http://www.opengeospatial.org/standards/requests/80>, 2011. Document 11-052r3.
- 44 Open Street Map Foundation. OpenStreetMap. URL: <https://www.openstreetmap.org/>.
- 45 Hoan Nguyen Mau Quoc and Danh Le Phuoc. An elastic and scalable spatiotemporal query processing for linked sensor data. In *11th International Conference on Semantic Systems (Semantics'15)*, pages 17–24, 2015. doi:10.1145/2814864.281486.
- 46 Amir Raza. Comparison of geospatial support in rdf stores: Evaluation for icos carbon portal metadata. *Master Thesis in Geographical Information Science*, 2019.
- 47 RDF Core Working Group. Resource Description Framework (RDF): Concepts and Abstract Syntax. <http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/>, February 2004.
- 48 Blake Regalia, Krzysztof Janowicz, and Grant McKenzie. Computing and querying strict, approximate, and metrically refined topological relations in linked geographic data. *Transactions in GIS*, 23(3):601–619, 2019. doi:10.1111/TGIS.12548.
- 49 Philippe Rigaux, Michel Scholl, and Agnes Voisard. *Spatial databases: with application to GIS*. Elsevier, 2002.
- 50 S2 Geometry. URL: <http://s2geometry.io/>.

- 51 Marina Schauffler. Testing the waters: Tracing the movement of PFAS into waterways and wildlife, 2023. URL: <https://themainemonitor.org/testing-the-waters-tracing-the-movement-of-pfas-into-waterways-and-wildlife/>.
- 52 Katrina Schweikert, David Kedrowski, Shirly Stephen, and Torsten Hahmann. SAWGraph Example Geospatial SPARQL Queries. Software, version 1.1, swbId: swb:1:dir:678ee78feb48f235c42bd5722e4c19f81f91f9dc (visited on 2025-07-30). URL: <https://github.com/SAWGraph/public/tree/main/UseCases/UC3-Tracing/UC3-CQ15/GIScience2025-queries>, doi:10.4230/artifacts.24220.
- 53 Chandan Sharma, Pierre Genevès, Nils Gesbert, and Nabil Layaïda. Schema-based query optimisation for graph databases. *arXiv preprint arXiv:2403.01863*, 2024. doi:10.48550/arXiv.2403.01863.
- 54 Cogan Shimizu, Shirly Stephen, Adrita Barua, Ling Cai, Antrea Christou, Kitty Currier, Abhilekha Dalal, Colby K Fisher, Pascal Hitzler, Krzysztof Janowicz, et al. The KnowWhere-Graph ontology. *Journal of Web Semantics*, page 100842, 2024.
- 55 Amila O. De Silva, James M. Armitage, Thomas A. Bruton, Clifton Dassuncao, Wendy Heiger-Bernays, Xindi C. Hu, Anna Kärrman, Barry Kelly, Carla Ng, Anna Robuck, Mei Sun, Thomas F. Webster, and Elsie M. Sunderland. PFAS exposure pathways for humans and wildlife: A synthesis of current knowledge and key gaps in understanding. *Environmental Toxicology and Chemistry*, 40:631–657, March 2021. doi:10.1002/etc.4935.
- 56 Shirly Stephen, Mitchell Faulk, Krzysztof Janowicz, Colby Fisher, Thomas Thelen, Rui Zhu, Pascal Hitzler, Cogan Shimizu, Kitty Currier, Mark Schildhauer, et al. The S2 hierarchical discrete global grid as a nexus for data representation, integration, and querying across geospatial knowledge graphs. *arXiv preprint*, 2024. arXiv:2410.14808.
- 57 Shirly Stephen, Torsten Hahmann, and David K. Kedrowski. The SAWGraph spatial ontology. URL: <https://raw.githubusercontent.com/SAWGraph/geospatial-kg/refs/heads/main/ontologies/sawgraph-spatial-ontology.ttl>.
- 58 Markus Stocker, Andy Seaborne, Abraham Bernstein, Christoph Kiefer, and Dave Reynolds. SPARQL basic graph pattern optimization using selectivity estimation. In *17th International Conference on World Wide Web (WWW 2008)*, pages 595–604. ACM, 2008. doi:10.1145/1367497.136757.
- 59 Kerry Taylor, Simon Cox, Krzysztof Janowicz, Maxime Lefrançois, Danh Le Phuoc, and Armin Haller. Semantic sensor network ontology. W3C recommendation, W3C, October 2017. URL: <https://www.w3.org/TR/2017/REC-vocab-ssn-20171019/>.
- 60 Kingsley Wiafe-Kwakye, Torsten Hahmann, and Kate Beard. An ontology design pattern for spatial and temporal aggregate data (STAD). In *13th Workshop on Ontology Design and Patterns (WOP 2022) at the 21st International Semantic Web Conference (ISWC 2022)*. CEUR-WS.org, 2022. URL: <https://ceur-ws.org/Vol-3352/pattern4.pdf>.
- 61 Marc Wick. GeoNames. URL: <https://www.geonames.org/>.
- 62 Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The FAIR guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.
- 63 Rui Zhu. Geospatial knowledge graphs. *arXiv preprint arXiv:2405.07664*, 2024. doi:10.48550/arXiv.2405.07664.

A Namespaces for ontologies and SPARQL queries

■ **Table 2** Ontology namespaces used for the queries in Section 5, the standard namespaces for RDF, RDFS, OWL, and XSD are omitted here.

PREFIX	Ontology namespace (URL)
coso:	http://w3id.org/coso/v1/contaminoso#
fio:	http://w3id.org/fio/v1/fio#
geo:	http://www.opengis.net/ont/geosparql#
hyf:	https://www.opengis.net/def/schema/hy_features/hyf/
kwg-ont:	http://stko-kwg.geog.ucsb.edu/lod/ontology/
kwgr:	http://stko-kwg.geog.ucsb.edu/lod/resource/
me_egad:	http://w3id.org/sawgraph/v1/me-egad#
naics:	http://w3id.org/fio/v1/naics#
nhdplusv2:	http://w3id.org/hyfo/v1/nhdplusv2#
qudt:	http://qudt.org/schema/qudt/
spatial:	http://purl.org/spatialai/spatial/spatial-full#

B Visualization of the Contaminant Tracing Results for the State of Maine

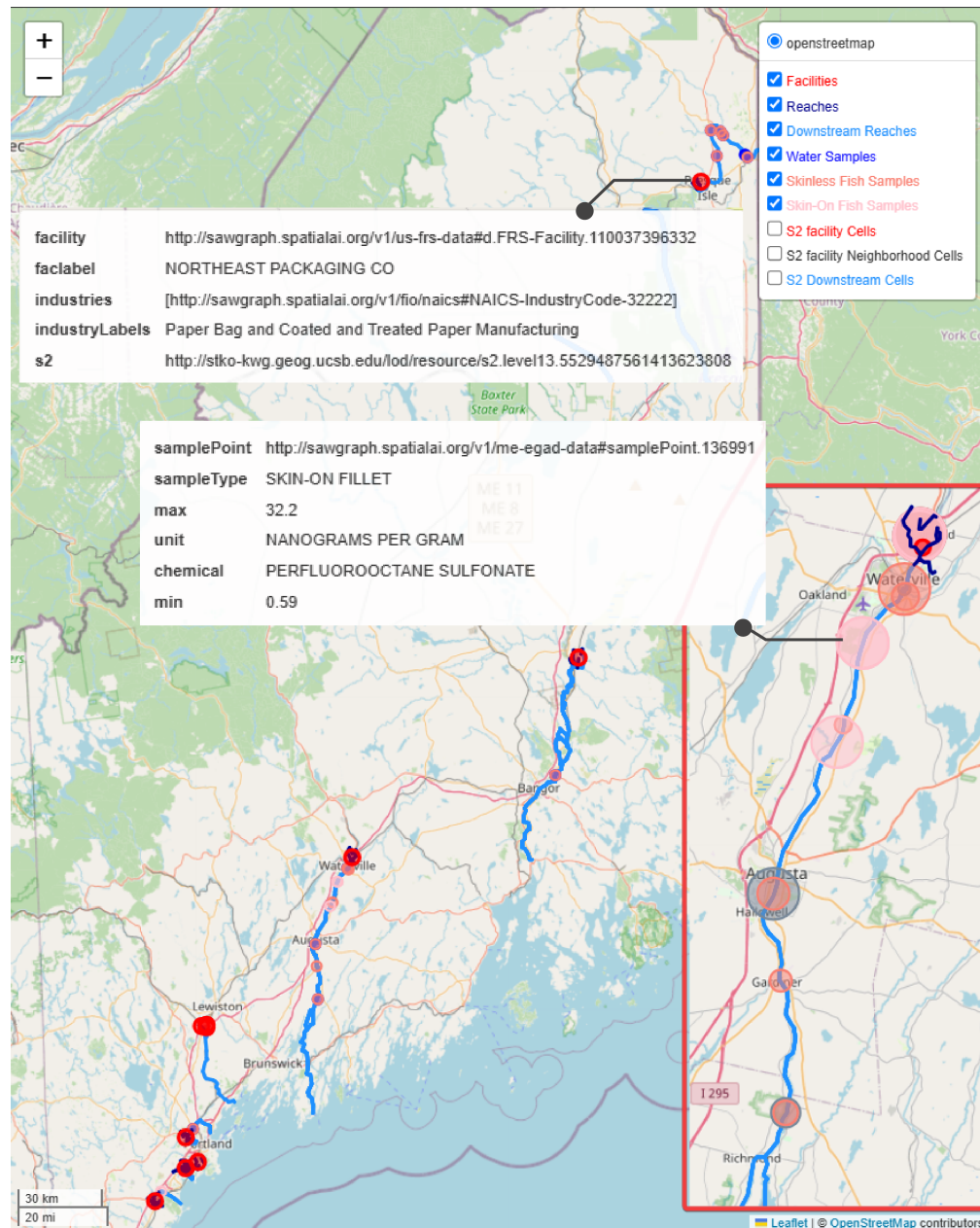


Figure 6 Screenshot of the interactive visualization of fish tissue and surface water sampling results downstream of ten paper manufacturing facilities in Maine. The inlay focuses on the results along the Kennebec River.

Analysis of Points of Interests Recommended for Leisure Walk Descriptions

Ehsan Hamzei ✉ 

University of Melbourne, Parkville, Australia

Thi Minh Hoai Bui ✉ 

University of Melbourne, Parkville, Australia

Martin Tomko ✉ 

University of Melbourne, Parkville, Australia

Stephan Winter ✉ 

University of Melbourne, Parkville, Australia

Abstract

Leisure walking is a physical activity where locomotion through a natural or even urban environment is the goal in itself, e.g., in pursuit of health and wellbeing. In contrast to destination-oriented walks that are focused on navigation efficiency (i.e., shortest or simplest walk from source to destination), leisure walks emphasize experiencing the environment, engaging in activities, and discovering places that may be off route, or intermediate destinations en-route, summarily called points of interest (POIs). POIs are key for recommending leisure walks, yet a detailed analysis of POIs in the context of leisure walking is missing in the literature. This study extracts and annotates POIs of leisure walking recommendations available in *WalkingMaps.com.au*, creating an annotated dataset to address this research gap and provide a first analysis of leisure walking descriptions. We classify POIs using the verbal description provided in the dataset, match them with data available in OpenStreetMap (OSM), and compare the POIs with nearby alternatives in OSM. Our analysis reveals thematic and spatial patterns in POI selection, offering a machine learning approach to model POI choices for leisure walks. We further evaluate the availability of rich data in OSM for future automated leisure walking recommendation. This study contributes to automated systems for recommending leisure walks, tailoring suggestions based on available information in the spatial open data, and presents an annotated dataset to facilitate future research in this field.

2012 ACM Subject Classification Information systems → Geographic information systems; Information systems → Location based services

Keywords and phrases leisure walks, points of interest, places, platial information

Digital Object Identifier 10.4230/LIPIcs.GIScience.2025.5

Supplementary Material *Software (Source code)*: <https://github.com/hamzeiehsan/leisure-walking-analysis> [8], archived at `swb:1:dir:cddb6d133e212246c2e458e4ea46f1358cd27927`

1 Introduction

Leisure walking is a physical activity where locomotion through a natural or even urban environment is the goal in itself, e.g., in pursuit of health and wellbeing. In contrast to the everyday walks, which are *destination-oriented* (where the path is merely a means of reaching the destination), leisure walks are *path-oriented* with a focus on the experience of the environment along the path [20]. In other words, in leisure walks the points of interest experienced at distance (off-route) or visited en-route as intermediate destinations, are as important or maybe even more important than the final destination, which may be identical with the start location.



© Ehsan Hamzei, Thi Minh Hoai Bui, Martin Tomko, and Stephan Winter;
licensed under Creative Commons License CC-BY 4.0

13th International Conference on Geographic Information Science (GIScience 2025).

Editors: Katarzyna Sila-Nowicka, Antoni Moore, David O'Sullivan, Benjamin Adams, and Mark Gahegan;
Article No. 5; pp. 5:1–5:16



Leibniz International Proceedings in Informatics
LIPIcs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Destination-oriented walks prioritize efficiency, enabling walk recommendations using optimal path algorithms (e.g., shortest or simplest paths) and spatial data typically stored in graph structures. Such walks can be externalized into verbal descriptions focused on efficient and effective navigation, with survey or route perspectives as the narrative strategy [25]. In contrast, leisure walk recommendations require platial information to suggest walks that focus on experiencing the environment [22, 23]. Leisure walk descriptions also demand a more intricate approach, moving beyond navigation, to offer rich descriptions of where to stop, what to see, and what activities to do, alongside narratives about places and their social and historical significance [23, 24].

Previous studies analyzing walk recommendations and descriptions provide insights into how people choose and communicate landmarks for navigation purposes [19, 10]. These studies are either limited to a specific walk [20] or a specific type of environment (e.g., natural landscape) [19], or they focus solely on route instructions [10] rather than the characteristics of points of interest (POIs) that fundamentally shape the experience of leisure walking. A detailed analysis of how people communicate descriptions of leisure walks and how they select relevant POIs is thus missing in the literature. This study aims to find insights and patterns of the characteristics of the chosen POIs in a leisure walking corpus.

The study addresses gaps in leisure walking recommendation research, particularly the lack of a dedicated dataset for POIs in the context of leisure walking. By collecting data from the *WalkingMaps* website¹, a sharing platform for leisure walking experiences in Victoria, Australia, we create an annotated dataset covering both leisure walk descriptions and POIs. To investigate why POIs are relevant for leisure walks, we classify the chosen POIs in the dataset and discuss their similarities and differences compared to nearby available POIs in OSM. Additionally, we conduct an evaluation of OSM data for leisure walking recommendations. Using a semi-automatic matching approach, we assess the availability of the selected POIs in OSM, contributing insights into the feasibility of automated leisure walk recommendations using open data.

This study hypothesizes that the thematic and spatial characteristics of POIs within a geographic extent can be used to identify patterns that describe which POIs are relevant for a leisure walking experience. To investigate this hypothesis, the following research questions must be addressed:

- What types of POIs are selected in leisure walking recommendations?
- To what extent is rich thematic and spatial information for the recommended POIs available in OSM?
- How can a machine learning model imitate the POI selection process for a given geographic area?

In short, the contributions of this study are:

- A leisure walk recommendation dataset, enriched with recommended POIs matched to OSM objects;
- A classification of the POIs in the dataset;
- A preliminary analysis of the availability of rich data in OSM for such POIs;
- A baseline prediction model for choosing relevant POIs for a leisure walk given a geographic extent.

¹ <https://walkingmaps.com.au/>

2 Related Works

POIs are defined as locations or objects that cartographers add to maps using cartographic symbols or labels to communicate relevant places [16]. Alternatively, POIs can be described as specific locations that individuals might find interesting or useful [9, 3]. Both definitions emphasize the importance of *relevance and interest* to the individual seeking POI recommendations. Consequently, POI identification is context-dependent and inherently subjective. For instance, POIs recommended to tourists may differ from those relevant to residents of a neighborhood, and the concept of POIs in urban analytics varies from that in mobility studies. Some research has narrowed the scope of POI to systematic definitions of specific types of POIs. For example, natural POIs have been defined in tourism and conservation management contexts using a structured rubric, with applicability demonstrated through examples from *OSM*, *iNaturalist*², and *Scenic-or-Not*³ data sources [9].

Categorizing POIs has been a major focus in POI-related research, using topic modeling in a semi-supervised manner to identify meaningful taxonomies for describing POIs in specific datasets (e.g., [5, 21, 6, 11]). These categorizations often result in POI classes relevant to specific domains, frequently focused on urban spaces due to the availability of rich datasets [21, 5, 11]. The Latent Dirichlet Allocation (LDA) topic modeling method is commonly used to identify dominant POI classes in textual datasets such as Foursquare⁴ and Yelp⁵ [11]. Typical POI types identified in these studies include *restaurants, cafes, and bars, shops and malls, public spaces such as parks and squares, museums, and religious and historic sites*. Additional classes depend on the dataset and method; for instance, [5] identified beach-related categories (e.g., beaches, piers, surf spots) due to their geographic focus, while [6] included businesses, transportation facilities, and government, health, and education-related POIs. However, these studies are predominantly focused on urban areas and do not specifically address a leisure walk context, where POIs are locations along a route, sometimes at the cost of few more steps from the walk, to spend a short time to view, visit, explore and interact [17].

Other common areas in POI-related research include predicting POIs based on previously selected POIs along a path [3, 27], efficiently storing and retrieving POIs [15], and evaluating the availability and quality of POI datasets [26]. These studies often provide general-purpose solutions for POI research and do not specifically address the unique challenges associated with domains such as leisure walks. For instance, predicting POIs for leisure walks involves distinct challenges: a walk may have a specific theme (e.g., visiting historic landmarks or bird habitats in rural areas), influencing what should be considered relevant as a POI. Additionally, POIs for leisure walks are not necessarily tourist attractions or places selected solely for navigational purposes but may include lesser-known places and objects discovered through personal experiences, often appealing especially to local residents [17]. Regarding data quality and availability, POIs associated with leisure walks often include a diverse set of less prominent places and objects, which means general-purpose studies may not adequately represent the specific data quality and availability conditions for these POIs.

This study addresses these research gaps by collecting and annotating a dataset specifically relevant to leisure walking POI research. We classify the collected POIs using a topic modeling approach and analyze the availability and quality of leisure walk POIs in OSM, examining

² <https://www.inaturalist.org/>

³ <https://scenicornot.datasciencelab.co.uk/>

⁴ <https://location.foursquare.com/products/places>

⁵ <https://www.yelp.com/dataset>

how the identified classes relate to data quality aspects. Finally, we present a baseline machine learning model to imitate human selectivity in choosing relevant POIs for leisure walks and discuss the challenges involved in developing such predictive models.

3 Leisure Walking Dataset

WalkingMaps is a publicly available service provided by *Victoria Walks Inc.*, a non-profit organization dedicated to promoting walkable communities in Australia. The platform allows users to explore and share leisure walks. Each recommended leisure walk includes several types of information: (1) the *walk*, represented as a linear geometry; (2) *POIs*, each represented as a point with a verbal description and optionally an image; and (3) a *verbal description* of the walk, highlighting the experience and providing navigational instructions. We utilized web scraping techniques to extract this information from the *WalkingMaps* website, resulting in a dataset of 386 leisure walks and 4392 POIs⁶. Detailed statistics derived from the verbal descriptions of the walks and POIs are presented in Table 1. The table shows that POI descriptions average about 23 words, sufficient to describe the POI and convey the rationale for the POI recommendation.

■ **Table 1** Statistics of collected textual data from *WalkingMaps*.

Item	Min	Median	Mean	Max
Walk description (word count)	7	130	181	540
Walk description (character count)	43	764	1062	3052
POI description (word count)	2	23	22	116
POI description (character count)	11	127	130	296

To further enrich the collected data, we developed a semi-automated approach to match the POIs with OSM objects. Using the Nominatim Geocoding API provided by OSM, we found that only 14.16% of the POI descriptions could be automatically matched to OSM data. This low geocoding rate is attributed to the rich verbal descriptions included in the POI data, well beyond simple names and feature types, making it challenging for the Geocoding API to interpret and locate the corresponding objects in OSM.

To achieve more accurate matching between POIs in the *WalkingMaps* dataset and OSM, we designed a simple annotation interface. This interface displays the verbal description of a POI, its location on the OSM map, and the ten OSM objects with the highest matching scores to the POI description, facilitating the interactive identification of relevant OSM objects. The matching score is calculated using the cosine similarity of embedding vectors derived from the POI verbal description and concatenated OSM key values for nearby OSM objects. These embeddings are generated using sentence transformers [18] with the *msmarco-distilbert-dot-v5* model. We selected this model, trained on the MS MARCO dataset [1], because its characteristics align closely with our problem of matching POI descriptions to OSM thematic representations. In the MS MARCO dataset, queries consist of keyword-based prompts (e.g., “largest river”) paired with natural language answers (e.g., “The Nile River is the longest river in the world at 4,132 miles”), which aligns to our task of mapping less structured, keyword-like OSM representations (e.g., “nature beach”) to rich textual descriptions of *WalkingMaps* POIs (e.g., “Byron Bay Main Beach is a walker’s paradise. Flat, hard sand and plenty of things to look at.”).

⁶ This dataset has been collected in February 2023.

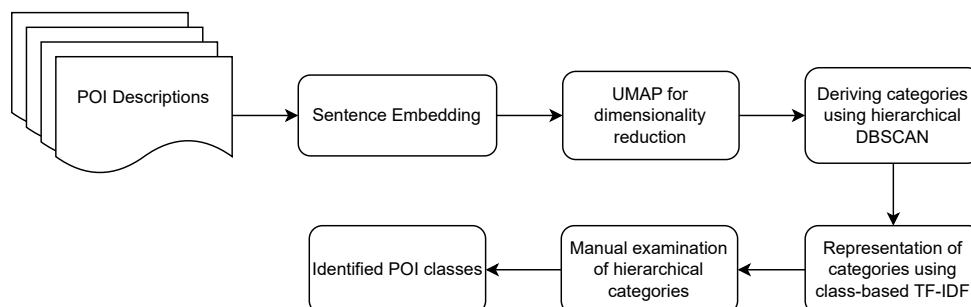
Using the annotation interface, annotators can select matches from the top ten suggestions or explore the map to identify other relevant objects in OSM. When a match is found, the annotator records the OSM identifier(s) alongside the corresponding POI in the dataset. Three scenarios were observed during the annotation process: (1) no match, when the POI is missing in OSM; (2) single match, when only one object matches the POI description; and (3) multiple matches, when several objects align with the POI description. The third scenario occurs when a POI description refers to an aggregate of similar objects (e.g., “there are a couple of ponds in the garden”) or a collection of spatial objects of different types (e.g., “a great picnic area and playground, complete with multiple BBQs, toilets, and plenty of play equipment”).

In total, 2385 POIs were matched with 3119 OSM objects. Due to multiple matches, the number of matched OSM objects exceeds the number of POIs in the leisure walking dataset. Among the 2385 matched POIs, 2022 are single matches, while 363 are multiple matches, associated with 1097 OSM objects. The remaining 2007 POIs could not be matched with any OSM objects at the time of annotation (i.e., May-June 2024).

To study the subjectivity involved in matching POIs to OSM objects, 5% of the dataset was independently annotated by another annotator. The results showed an inter-annotator agreement of 81.2%, with 407 out of 497 POIs in complete agreement between the two annotators. When considering partial agreements cases, where multiple matches for a POI resulted in overlapping lists of matches between the two annotators, the agreement increased to 88.1%, covering 438 out of 497 POIs.

4 POI Classification

In the first step, we use topic modeling to categorize POIs based on their verbal descriptions (see Figure 1). To ensure the topics are derived from the intrinsic characteristics of the POIs (i.e., what the POIs are) rather than their geographic locations (i.e., where the POIs are located), we pre-process the text to remove names of common places such as suburbs, local government areas, towns, and cities. For topic modeling, we apply the BERTopic workflow [7], which involves transforming descriptions into sentence embeddings, reducing the dimensionality of the embeddings using Uniform Manifold Approximation and Projection (UMAP) [14], clustering the reduced vectors with the Hierarchical DBSCAN method [2] to minimize noise, and creating bag-of-words representations for each cluster. The clusters are then characterized by top keywords identified using topic class-based Term Frequency-Inverse Document Frequency (TF-IDF). This workflow results in a detailed, hierarchically organized list of topics, each represented by a set of keywords.



■ **Figure 1** Classification workflow.

To identify classes based on the identified topics, we manually examine each topic within the hierarchy. We verify whether the top-10 keywords for each topic represent a particular type of places (i.e., well-formed topics). We then interpret the hierarchy of these well-formed topics to label and classify them. The proposed classification of POIs, based on the 75 identified topics, includes four classes further divided into ten sub-classes (209 POIs belonging to five malformed topics are not classified, and labeled as unknown). The four classes are:

- **Nature-related POIs:** This class includes 22 identified topics related to nature, describing natural features such as mountains, bays, beaches, and waterfalls, as well as habitats for flora and fauna (e.g., “the native grass trees are particularly striking”). In the dataset, 1478 POIs are classified as nature-related, 745 as natural features and 733 as habitats for flora and fauna. Out of 386 leisure walks, 338 include at least one nature-related POI, and 125 walks have the majority of POIs (at least 50%) belonging to this class. This indicates that nature-related POIs are the dominant class in terms of both the number of POIs and the number of leisure walks that include these POIs. The definition of nature-related POIs extends beyond the previous definition by [9] and encompasses places or regions with fuzzy boundaries that describe, for example, habitats for flora and fauna.
- **Activity-related POIs:** These POIs are described in terms of their functional roles and what activities they offer, which is the primary reason for their recommendation. For example, “gather your friends and start training for the WNBL or NBL by shooting some hoops at Braybrook Park” (the (women’s) national basketball league). The four major subclasses of this category are places that support sport-related activities (212 POIs), aquatic activities (227 POIs), picnic and camping (482 POIs), and food and beverages (167 POIs). This class includes 1088 POIs in the dataset, with 315 leisure walks having at least one activity-related POI and 63 walks having the majority of their POIs belonging to this class.
- **Society-related POIs:** These POIs are man-made features and objects that people found interesting and recommended for investigation during leisure walks in the neighborhood. For example, “established by Aboriginal artist Lin Onus, this was a social and political meeting place during the 1960s for young people influenced by the Black Power Movement.” This class includes two major subclasses: (1) human-made landmarks (e.g., hospitals, hotels, and law courts), and (2) places and objects with artistic, historical, and cultural significance (e.g., native people establishments, historic gold mines, and murals). This class includes 1070 POIs in our dataset, with 376 human-made landmarks and 694 POIs with artistic, historical, and cultural significance. In total, 256 leisure walks include POIs of this class, with 64 walks having the majority of their POIs belonging to this class.
- **Transport-related POIs:** This class includes two major subclasses: (1) trails, paths, streets, canals, and bridges that are described as POIs due to the atmosphere and experiences they afford (e.g., “the Boardwalk is fantastic and makes for easy walking”), and (2) transport-related facilities (i.e., lines and stations) that aid in planning commuting to and from the leisure walk area (e.g., “start at Ringwood Railway Station fronting Maroondah Highway”). This class includes a total of 547 POIs, with 472 related to trails, paths, and canals, and 75 describing lines and stations. In total, 238 leisure walks include POIs of this class, with twelve walks having the majority of their POIs belonging to this class.

5 POIs and Data Quality of OSM

As discussed in Section 3, we matched the POIs from the WalkingMaps dataset to OSM objects wherever possible. Here, we present detailed results on the matching status (no match, single match, multiple matches) of the POI classes and subclasses. We further discuss how these classes and subclasses are described using OSM tags (i.e., key/value pairs).

Table 2 shows the matching status of POIs based on their class and subclass. The results reveal that most unmatched POIs belong to the nature-related class, specifically to habitats of flora and fauna (552 POIs, roughly 75%). One reason for this missing data in OSM is the complexity of mapping habitat regions due to their fuzzy and possibly unknown boundaries. Table 2 shows that activity-related POIs are mostly found in OSM, but with a considerable number of POIs in this category matched to multiple OSM objects. This is because their POI descriptions may include several features (e.g., multiple shops in a neighborhood, or BBQ facilities and seats in a park) or a single feature conceptualized as multiple entities in OSM (e.g., a sports complex with multiple buildings). Society-related and transport-related POIs are also mostly found in OSM as well. The unmatched POIs for these classes are due to (1) places that no longer exist (e.g., an former shop that has changed, no longer exists, or was demolished), (2) places or features that have not been mapped, mostly trails and small artistic objects (e.g., sculptures and murals), and (3) a difficulty in identification due to a lack of thematic information for OSM objects related to these classes.

■ **Table 2** Matching status for each class and subclass.

Class	Subclass	Matching status
Nature-related	flora and fauna	no match (552), single match (152), multiple matches (29)
Nature-related	natural landmarks	no match (322), single match (357), multiple matches (66)
Activity-related	picnic and camping	no match (158), single match (257), multiple matches (67)
Activity-related	aquatic	no match (80), single match (112), multiple matches (35)
Activity-related	sport	no match (76), single match (108), multiple matches (28)
Activity-related	food and beverage	no match (54), single match (105), multiple matches (8)
Society-related	human-made landmarks	no match (122), single match (238), multiple matches (16)
Society-related	art, history and culture	no match (351), single match (310), multiple matches (33)
Transport-related	trails, paths and canals	no match (165), single match (258), multiple matches (49)
Transport-related	lines and stations	no match (27), single match (44), multiple matches (4)

Table 3 shows the top OSM tags for each class and subclass of the POIs matched to the OSM database. The results indicate that for the nature-related class, the dominant keys are natural, leisure, foot, and highway, providing information about the type of places

and their accessibility. Activity-related POIs are mainly described using the leisure key with values such as playground, pitch, and park, with common tags such as the building tag also being popular. Society-related POIs are described primarily using the building tag, along with specific tags for places of worship and the tourism key. Other popular tags for society-related POIs, not listed in the table, include `tourism:[museum, hotel]`, `amenity:[school, post_office]`, and `historic:memorial`, highlighting their cultural and historical significance. The transport-related class includes popular tags describing types, such as `bridge:yes`, `highway:cycleway` for the trails, paths, and canals subclass, and `railway:station`, `train:yes` for lines and stations. As shown in the table, the number of frequent tags, except for common tags like `building:yes`, is much smaller than the number of actual matched POIs for each class and subclass. This discrepancy is due to either the POIs belonging to diverse categories or the OSM records lacking sufficient key-value information to describe them (i.e., 1078 OSM objects in this experiment only had one tag, either name or type, without any other thematic information available).

■ **Table 3** Popular OSM key/values for each class and subclass.

Class	Subclass	Most frequent OSM tags (count)
Nature-related	flora and fauna	leisure:park (30), natural:water (29), highway:footway (19)
Nature-related	natural landmarks	foot:yes (102), highway:cycleway (102), highway:path (70)
Activity-related	picnic and camping	leisure:playground (98), access:yes (58), leisure:park (43)
Activity-related	aquatic	man_made:pier (19), building:yes (13), leisure:slipway (11)
Activity-related	sport	leisure:pitch (63), leisure:park (21), building:yes (18)
Activity-related	food and beverage	building:yes (21), amenity:restaurant (19), amenity:cafe (18)
Society-related	human-made landmarks	building:yes (56), religion:christian (23), amenity:place_of_worship (22)
Society-related	art, history and culture	building:yes (57), addr:state:VIC (41), tourism:artwork (31)
Transport-related	trails, paths and canals	layer:1 (92), bridge:yes (90), highway:cycleway (67)
Transport-related	lines and stations	railway:station (15), train:yes (11), railway:miniature (10)

The relationship between the identified classes and subclasses with OSM tags (key-value pairs) can be numerically described using Cramér's V categorical association [4]. Table 4 shows the measured associations between OSM keys, OSM values, and OSM key-value pairs (i.e., tags) with classes, subclasses, and topics. It indicates that key-value pairs have a high association value with classes (0.73) and subclasses (0.70). With both key and value available, we can predict the corresponding class and subclass. This number is lower for individual keys and values, especially for keys, which are often general – e.g., the *leisure* key can have values such as park, fishing, garden, or pitch, describing POIs belonging to different classes and subclasses in our classification.

■ **Table 4** Cramér’s V categorical association between class/subclass/topic.

	OSM key	OSM value	OSM key-value
Class	0.43	0.68	0.73
Subclass	0.34	0.67	0.70
Topic	0.20	0.61	0.63

To demonstrate the impact of low-quality thematic information on classifying POIs using OSM tags, we trained two baseline predictive models. The first model uses a Ridge Classification approach with TF-IDF vectors of all concatenated key-value tags available in OSM as input features. The second model is a Gradient Boosting Classification model trained with the presence of popular key values as binary features (defined by a hard threshold of at least 25 counts, resulting in 472 keys). Both classifiers were tested using 10-fold cross-validation, showing similar results in predicting POI classes using OSM thematic information (see Table 5). Removing OSM matches that only have a name and no further descriptive tags significantly improves the accuracy of both models from about 0.54 to 0.72. This indicates that even with few available tags, we can identify POI types; however, a significant number of matched POIs do not have such descriptive tags available. The *SHapley Additive exPlanations* (SHAP) values [12] of the Ridge model and the feature importance measures of the Gradient Boosting model highlight the importance of features such as *amenity*, *playground*, *viewpoint*, *natural*, *waterfall*, *building*, *garden*, and *office* in the thematic descriptions to predict the correct POI classes.

■ **Table 5** Accuracy, weighted precision, recall and F-score values of the baseline predictive models.

	Accuracy	Precision	Recall	F-score
Ridge Classifier	0.54	0.6	0.54	0.53
Gradient Boosting Classifier	0.55	0.58	0.55	0.53

6 POI Selection

In this section, we analyze the selectivity involved in choosing a POI for a leisure walk and test a baseline machine learning (ML) method to predict whether a POI candidate is suitable for a leisure walk (i.e., a binary classifier). We already have the selected POIs in the dataset (the matched POIs to the OSM database with at least one tag other than *name*, i.e., 2367 OSM objects). To define alternative POI choices not recommended for leisure walks, we selected all other OSM objects within 200 meters of the actual leisure walk POIs (extending the buffer zone to 1000 meters if no alternative POI candidate was found) and filtered them to spatial objects containing one or more of the following OSM keys (with any value) in their tags: *amenity*, *shop*, *railway*, *bridge*, *club*, *building*, *historic*, *tourism*, *place*, *waterway*, *landuse*, *leisure*, *natural*, *office*, *boundary*, *highway*, *man_made*. These keys were selected based on the top frequent tags observed in OSM objects matched with the leisure walk POIs.

The list of POI candidates that are not recommended by people within their leisure walk descriptions contains 183906 spatial objects with at least one tag other than *name*. The number of not recommended POIs is more than 77 times larger compared to the recommended POIs in our dataset, highlighting the selectivity and challenge of POI selection. The random

baseline theoretically results in 1.3% accuracy, meaning only one out of 77 suggestions is a POI recommended by people for leisure walks. Another challenge is related to the sparsity of the data; some not-recommended POIs could be relevant for (other) leisure walks, but due to the relatively small number of recommended walks in the dataset, their relevance is unknown and not being captured.

To test how a baseline ML model can imitate human selectivity in recommending a POI for a leisure walk, we trained a binary classifier model and select the best performing model using 10-fold cross-validation. This classifier includes an ensemble of two classifiers focused on the available thematic and spatial features of the OSM objects, respectively. Both classifiers are trained with over-sampled data using the SMOTE technique to better model this highly imbalanced dataset, and tested with an imbalanced unseen dataset. The test dataset includes 10% of the whole dataset, randomly selected and stratified based on the binary outputs to ensure it includes both positive and negative cases.

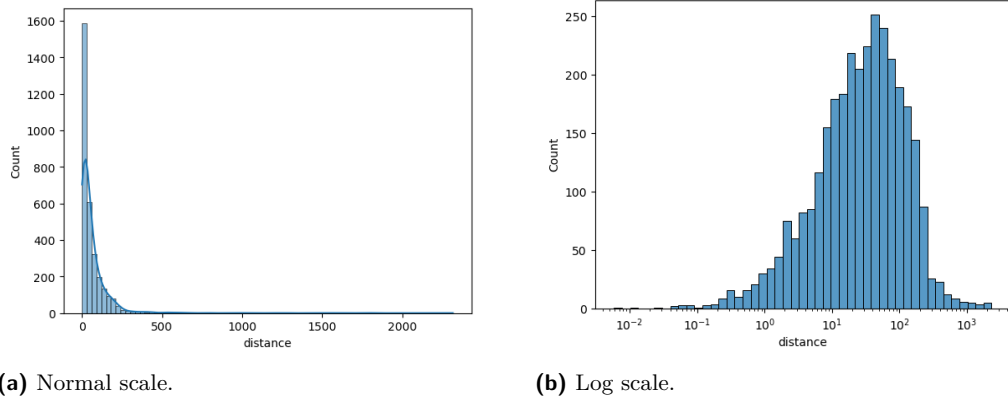
The thematic classifier is trained using the Ridge Classification method with features being TF-IDF vectors of textual descriptions generated by concatenating OSM key-value pairs. Table 6 shows the confusion matrix of the test dataset for predicting whether a candidate is a relevant POI in the leisure walk context. This classifier has an overall accuracy of 0.90, precision of 0.10, recall of 0.82, and ROC AUC of 0.86. These results indicate that while the model performs much better than a random classifier and has a high recall, it still struggles with precise predictions. In other words, out of 10 predicted POIs, only one was actually recommended by people, while most of the recommended POIs are predicted correctly (i.e., 82.3% recall). The low precision is primarily attributed to the subjectivity inherent in POI selection and not necessarily that suggested POIs are irrelevant to the context of leisure walk. Although several POIs may be relevant and useful for providing a leisure walk experience, only a few are suggested by users, influenced by their personal preferences and experiences. However, a much larger and more diverse dataset is required to minimize the impact of subjectivity in POI selection task.

■ **Table 6** Confusion matrix predicting relevant POIs using thematic information.

	Not POI (predicted)	POI (predicted)
Not POI (actual)	16597	1794
POI (actual)	42	195

The spatial classifier is built based on the Euclidean distance from POI candidates belonging to the leisure walk, the area of POI candidates (0 for point-based and linear geometries), and their length (0 for point-based geometries). The histogram of distances from the recommended POIs to their leisure walks is shown in Figure 3. This figure demonstrates that the thresholds used for creating the list of the not-recommended POI candidates are within realistic distance ranges for the actual recommended POIs. Thus, we are not creating a single metric that makes the prediction obvious or trivial for the model. These three spatial features are used to train a Gradient Boosting classifier, and the confusion matrix of the test data is shown in Table 7. The results show slightly worse performance compared to the thematic classifier but are still much better than a random classifier. The accuracy is 0.82, precision is 0.06, recall is 0.84, and ROC AUC is 0.83.

When we ensemble the thematic and spatial classifiers, we observe the complementary role of these features through the improvement in prediction precision. The ensemble model predicts a POI candidate as relevant for leisure walks if both thematic and spatial classifiers



■ **Figure 3** Euclidean distance between the leisure walk path and matched OSM objects.

■ **Table 7** Confusion matrix for predicting recommended POIs using spatial information.

	Not POI (predicted)	POI (predicted)
Not POI (actual)	15067	3324
POI (actual)	39	198

predict it as relevant. The results of this ensemble model are shown in Table 8. This model achieves an accuracy of 0.98, precision of 0.34, recall of 0.69, and ROC AUC of 0.84. The significant improvement in precision, from one correct prediction out of ten predicted POIs for the thematic model to one correct prediction out of three predictions in the ensemble model, demonstrates the complementary role of thematic and spatial features in predicting the relevance of a POI candidate for leisure walks using OSM data.

■ **Table 8** Confusion matrix for predicting recommended POIs using spatial and thematic information.

	Not POI (predicted)	POI (predicted)
Not POI (actual)	18072	319
POI (actual)	74	163

7 Discussion and Conclusion

In this paper, we introduced a new dataset collected from the *WalkingMaps* website, which includes verbal descriptions of leisure walks, geometric representations of the walks, and a set of POIs with their point-based representations and verbal descriptions for each leisure walk. We further enriched the POIs by matching them with OSM objects using a semi-automated approach and classified each POI using a topic modeling based on their verbal descriptions. Our proposed classification includes four top-level classes: nature-related, activity-related, society-related, and transport-related POIs. The classification further breaks down into ten subclasses: habitats of flora and fauna, natural landmarks, places that offer food/beverages or sport/aquatic activities, places related to picnic and camping, human-made landmarks, places with artistic, historical, or cultural significance, transport-related places such as trails, paths, and canals, and finally lines and stations.

Next, we discussed the availability of OSM data for the POIs recommended in leisure walks. We observed that only 14.16% of the descriptions can be geocoded automatically. Even with manual inspection to find data in OSM, only 54.3% of POIs can be found in OSM database. During the annotation process, we noticed that several POI descriptions need to be matched with multiple OSM objects, either because these descriptions describe multiple objects at once (e.g., BBQ and playground in a park) or because the single object described is a complex entity modeled with multiple OSM objects (e.g., sports complexes that include several buildings). We noticed that several POI descriptions included sensory details (e.g., visual, auditory, olfactory, or tactile), which are highly relevant and useful in the context of leisure walking but are often ephemeral and context-dependent. These rich, human-centric descriptions were challenging to match with OSM records, as OSM often provides only rudimentary thematic information compared to the richness of these narratives. [13] conducted a detailed analysis of landscape and place descriptions incorporating sensory information; however, in the context of leisure walks, further research is needed to analyse these complex POI descriptions.

We further discussed the relationship between data availability and the identified classes, noting that roughly 75% of the POIs related to habitats of flora and fauna are missing in the OSM database. Activity-related POIs have more matches to multiple OSM objects compared to others, as their descriptions often include multiple activities offered in an area (i.e., multiple objects described in a description) or places with multiple buildings or compartments, modeled as multiple objects in OSM. Most activity and society-related POIs can be found in OSM, with missing ones mainly related to places that have changed use or no longer exist (e.g., a cafe or restaurant described in a leisure walk that now changed to another entity). The main reason for transport-related POIs not being matched to OSM objects is that the POI descriptions only describe part of a mapped path or trail, while OSM captures the whole path or trail. These parts of the paths in these cases are often important due to the views they offer or the atmosphere and vibe of walking in there (e.g., *“follow the gravel path down towards the falls for a spectacular view”*).

Finally, we study whether we can automate the POI selection for leisure walks, given the walk area, geometry, and a set of POI candidates (both the recommended POIs in the dataset and a large set of POI candidates from the OSM database). We trained an ensemble model that utilizes thematic information from OSM objects (all available tags) and their spatial features (i.e., distance to walk, area, and length). The results show the complementary role of spatial and thematic features, with the ensemble model significantly outperforming individual spatial and thematic models, improving from one correct guess out of 10 suggestions to one correct guess out of three suggestions. We also highlight the challenges of POI selection tasks due to subjectivity in the process, data sparsity (i.e., 2367 matched POIs), and highly imbalanced train/test datasets (one recommended POI for 77 not-recommended POI candidates).

This study can be further extended, by refining the methodology for classifying POIs for leisure walks and by further verifying our findings against leisure walk POIs from other datasets and geographies. Our focus in this paper is to provide baseline methods for selecting POIs and demonstrate these on the Leisure Walks dataset, to discuss its coverage and limitations. Using more sophisticated machine learning methods, we may expect improvements in the POI selection task. This task can be reformulated in other ways, such as predicting new POIs given a path and previous POIs. While beyond the immediate scope of this study, the dataset presented here can be used to develop and test such approaches. The data also have potential applications outside the leisure walking

context, such as studying and improving the automatic identification of OSM objects based on verbal descriptions, beyond the usual geocoding task, as here we consider multiple matches for a single description. The task here also differs conceptually from geocoding since the point-based location is already available and provided in the dataset. Instead of finding the place/object's location, we aim to find the OSM object(s) (or spatial entities in any other spatial database) related to what is verbally and geometrically described by a person. As described in the paper, the results of such matching may yield zero, one, or multiple objects depending on data availability and how objects are conceptualized within the database (e.g., OSM). For example, a sports complex may be represented as multiple areal features in OSM, yet described as a single entity in the verbal description.

In leisure walk descriptions research, it is essential to differentiate between two types of POIs: those where one can visit, investigate, and engage in activities, and those that serve as locations to view other objects of interest (e.g., a view from a hill to a distant mountain). Leisure walking, and the inclusion of POIs in walk descriptions demands further conceptual research to investigate the role of such POIs and why they are recommended by the authors of descriptions. This includes examining the purpose of a POI: is it recommended to aid navigation, for its functional role, or as a focus of attention for interest and enjoyment during the walk.

References

- 1 Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. MS MARCO: A Human Generated Machine Reading COmprehension Dataset, 2018. [arXiv:1611.09268](https://arxiv.org/abs/1611.09268).
- 2 Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. Density-based clustering based on hierarchical density estimates. In Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu, editors, *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.
- 3 Buru Chang, Yonggyu Park, Donghyeon Park, Seongsoon Kim, and Jaewoo Kang. Content-aware hierarchical point-of-interest embedding model for successive POI recommendation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI'18*, pages 3301–3307. AAAI Press, 2018. doi:10.24963/IJCAI.2018/458.
- 4 Harald Cramér. *Mathematical Methods of Statistics*. Princeton University Press, 1946.
- 5 Song Gao, Krzysztof Janowicz, and Helen Couclelis. Extracting urban functional regions from points of interest and human activities on location-based social networks. *Transactions in GIS*, 21(3):446–467, 2017. doi:10.1111/TGIS.12289.
- 6 Yunfan Gao, Yun Xiong, Siqi Wang, and Haofen Wang. Geobert: pre-training geospatial representation learning on point-of-interest. *Applied Sciences*, 12(24):12942, 2022.
- 7 Maarten Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*, 2022. doi:10.48550/arXiv.2203.05794.
- 8 Ehsan Hamzei, Thi Minh Hoai Bui, Martin Tomko, and Stephan Winter. hamzeiehsan/leisure-walking-analysis. Software, swbId: swb:1:dir:cddb6d133e212246c2e458e4ea46f1358cd27927 (visited on 2025-07-30). URL: <https://github.com/hamzeiehsan/leisure-walking-analysis>, doi:10.4230/artifacts.24214.
- 9 Charlie Hewitt, SD Sabbata, A Ballatore, S Cavazzi, and Nicholas Tate. Defining natural points of interest. In *29th Annual GIS Research UK Conference (GISRUK)*. Cardiff, Wales, UK (Online), 2021.
- 10 Olga Koblet and Ross S. Purves. From online texts to landscape character assessment: Collecting and analysing first-person landscape perception computationally. *Landscape and Urban Planning*, 197:103757, 2020. doi:10.1016/j.landurbplan.2020.103757.

- 11 Kang Liu, Ling Yin, Feng Lu, and Naixia Mou. Visualizing and exploring POI configurations of urban regions on POI-type semantic space. *Cities*, 99:102610, 2020. doi:10.1016/j.cities.2020.102610.
- 12 Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.
- 13 Nora Fagerholm Manuel F. Baer, Flurina Wartmann and Ross S. Purves. Extracting sensory experiences and cultural ecosystem services from actively crowdsourced descriptions of everyday lived landscapes. *Ecosystems and People*, 20(1):2331761, 2024. doi:10.1080/26395916.2024.2331761.
- 14 L. McInnes, J. Healy, and J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints*, February 2018. arXiv:1802.03426.
- 15 Grant McKenzie and Krzysztof Janowicz. OpenPOI: An open place of interest platform (Short paper). In *10th International Conference on Geographic Information Science (GIScience 2018)*. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2018. doi:10.4230/LIPIcs.GISCIENCE.2018.47.
- 16 Achilleas Psyllidis, Song Gao, Yingjie Hu, Eun-Kyeong Kim, Grant McKenzie, Ross Purves, May Yuan, and Clio Andris. Points of Interest (POI): A commentary on the state of the art, challenges, and prospects for the future. *Computational Urban Science*, 2(1):20, 2022.
- 17 Daniele Quercia, Rossano Schifanella, and Luca Maria Aiello. The shortest path to happiness: Recommending beautiful, quiet, and happy routes in the city. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, pages 116–125, 2014. doi:10.1145/2631775.2631799.
- 18 Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, November 2019. doi:10.18653/V1/D19-1410.
- 19 Tiina Sarjakoski, Pyry Kettunen, Hanna-Marika Halkosaari, Mari Laakso, Mikko Rönneberg, Hanna Stigmar, and Tapani Sarjakoski. Landmarks and a hiking ontology to support wayfinding in a national park during different seasons. In Martin Raubal, David M. Mark, and Andrew U. Frank, editors, *Cognitive and Linguistic Aspects of Geographic Space: New Perspectives on Geographic Information Research*, pages 99–119. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. doi:10.1007/978-3-642-34359-9_6.
- 20 Nancy Stevenson and Helen Farrell. Taking a hike: exploring leisure walkers embodied experiences. *Social & Cultural Geography*, 19(4):429–447, 2018. doi:10.1080/14649365.2017.1280615.
- 21 Demi van Weerdenburg, Simon Scheider, Benjamin Adams, Bas Spierings, and Egbert van der Zee. Where to go and what to do: Extracting leisure activity potentials from web data on urban space. *Computers, Environment and Urban Systems*, 73:143–156, 2019. doi:10.1016/J.COMPENVURBSYS.2018.09.005.
- 22 James Williams, Stefano Cavazzi, James Pinchin, Adrian Hazzard, Gary Priestnall, and Andrea Ballatore. Context for leisure walking routes: A vision for a spatial-platial approach. In *Spatial Data Science Symposium 2022 Short Paper Proceedings*, 2022. doi:10.25436/E20W2J.
- 23 James Williams, James Pinchin, Adrian Hazzard, Gary Priestnall, Stefano Cavazzi, and Andrea Ballatore. Emerging platial narratives and themes from a leisure walking study. In *Proceedings of the 4th International Symposium on Platial Information Science (PLATIAL’23)*, pages 23–28, 2023.
- 24 James Williams, James Pinchin, Adrian Hazzard, Gary Priestnall, Stefano Cavazzi, and Andrea Ballatore. Walkgis: Exploring platial analysis of leisure walks via linked video narratives. In *31st Annual Geographical Information Science Research UK Conference (GISRUK)*, 2023.

- 25 Stephan Winter, Ehsan Hamzei, Nico Van de Weghe, and Kristien Ooms. A graph representation for verbal indoor route descriptions. In Sarah Creem-Regehr, Johannes Schöning, and Alexander Klippel, editors, *Spatial Cognition XI*, pages 77–91, Cham, 2018. Springer International Publishing. doi:10.1007/978-3-319-96385-3_6.
- 26 Lih Wei Yeow, Raymond Low, Yu Xiang Tan, and Lynette Cheah. Point-of-Interest (POI) Data Validation Methods: An Urban Case Study. *ISPRS International Journal of Geo-Information*, 10(11), 2021. doi:10.3390/ijgi10110735.
- 27 Kangzhi Zhao, Yong Zhang, Hongzhi Yin, Jin Wang, Kai Zheng, Xiaofang Zhou, and Chunxiao Xing. Discovering subsequence patterns for next POI recommendation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3216–3222, 2021.

A Frequent Words by POI Class

■ **Table 9** Most frequent OSM key/values for each class and subclass.

Class	Most frequent OSM tags (count)
Nature-related	view (281), tree (224), bird (183), lake (182), river (165), beach (161), creek (154), path (148), see (130), track (116)
Activity-related	park (307), playground (208), picnic (164), area (151), BBQ (120), spot (102), river (101), great (95), place (86), club (83)
Society-related	garden (147), building (143), built (124), centre (117), art (114), church (103), street (81), memorial (78), community (77), school (77)
Transport-related	bridge (232), track (157), path (107), station (89), walking (83), railway (66), start (57), street (51), boardwalk (49), park (47)

MODAP: A Multi-City Open Data & Analytics Platform for Micromobility Research

Grant McKenzie   

Platialis Analysis Lab, McGill University, Montréal, Canada

Abstract

Over the past decade, micromobility services, particularly electric vehicles for personal short-distance trips, have experienced significant growth. Major cities around the world now host extensive fleets of vehicles available for short-term public rental. While previous research has examined usage patterns within and between a few select cities, large, open, and publicly accessible data sets for analyzing mobility across multiple cities are extremely limited. I have collected, curated, and aggregated over twenty million e-scooter and e-bicycle trips across five major cities and are openly releasing aggregated data for use by mobility and sustainable transport researchers, urban planners, and policymakers. To accompany these data, I developed MODAP (Micromobility Open Data & Analytics Platform), a geovisual analytics tool that empowers researchers to explore the temporal and regional patterns of e-mobility trips within our open data set and download the data for offline analysis. My objective is to foster further research into city-scale mobility patterns and to equip researchers, community members, and policymakers with the necessary tools to conduct this work.

2012 ACM Subject Classification Information systems → Geographic information systems; Human-centered computing → Geographic visualization

Keywords and phrases open data, mobility, geovisualization, micromobility

Digital Object Identifier 10.4230/LIPIcs.GIScience.2025.6

Supplementary Material *Software (Source Code)*: <https://github.com/grantdmckenzie/modap> [14] archived at `swh:1:dir:a48742b027327af1a11426dcff7ddf77efa10fb9`

1 Introduction

Analysis of urban mobility, namely investigating how people move through cities, is important for a wide range of applications such as tracking the spread of diseases, designing equitable and accessible cities, and mitigating the impacts of climate change. As urban populations continue to grow [27], gaining access to real urban mobility data has become increasingly important for policymakers, city planners and researchers [22]. While the curation of large-scale urban mobility data sets is expanding, access often comes with significant challenges. Traditional data sources, such as travel surveys, are costly to collect, while alternative sources, such as ride-hailing or social media check-in data, are typically siloed by private companies making them inaccessible to the public, researchers, and even municipal transport agencies. In recent years, data sharing has become even more restricted due to the threat of these data being used to train proprietary foundation models.

Urban transportation has undergone a shift over the past decade with the commercialization of existing modes of transport (e.g., shared bicycles) and the emergence of new micromobility options such as e-scooters. Shared micromobility systems, operated by private companies, have been deployed in hundreds of cities worldwide, offering fleets of short-term rental vehicles ranging from a few dozen to several thousand per city. Due to regulatory efforts taken by many municipalities, micromobility operators are often required to provide publicly accessible application programming interfaces (APIs) that report the real-time locations of available vehicles. While originally intended for regulatory oversight and safety compliance, these APIs have also enabled third-party integration, such as embedding them into navigation services like Google Maps.



© Grant McKenzie;
licensed under Creative Commons License CC-BY 4.0

13th International Conference on Geographic Information Science (GIScience 2025).

Editors: Katarzyna Sila-Nowicka, Antoni Moore, David O'Sullivan, Benjamin Adams, and Mark Gahegan;
Article No. 6; pp. 6:1–6:14



Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Today, shared micromobility has become an integrated component of urban transportation ecosystems, with over 172 millions trips having taken place in North America last year alone [21]. Access to trip-level micromobility data is invaluable for understanding mobility behavior, optimizing transportation networks, and informing policy decisions. In this work, I introduce an open, multi-city data set of micromobility trips, detailing the data collection process and limitations of my methodology. Additionally, I present a geovisual analytics platform that encourages users to interactively explore the data through a web-based interface. This platform enables users, including those without technical expertise, to visualize spatial patterns in trip distributions, temporal variations, and differences across micromobility modes, providing an accessible tool for urban mobility analysis. In more explicit terms, the objectives of this work are as follows.

1. To collect, clean, curate, and publish an open mobility data set of e-scooters and e-bicycle trips in five major cities around the world.
2. To develop a web platform to both serve the data and provide exploratory geovisual analytics functionality with the goal of democratizing data analytics and empowering those with limited ability or capacity to analyze the data offline.
3. To demonstrate the utility of these data through a showcase of several exploratory mobility analyses.

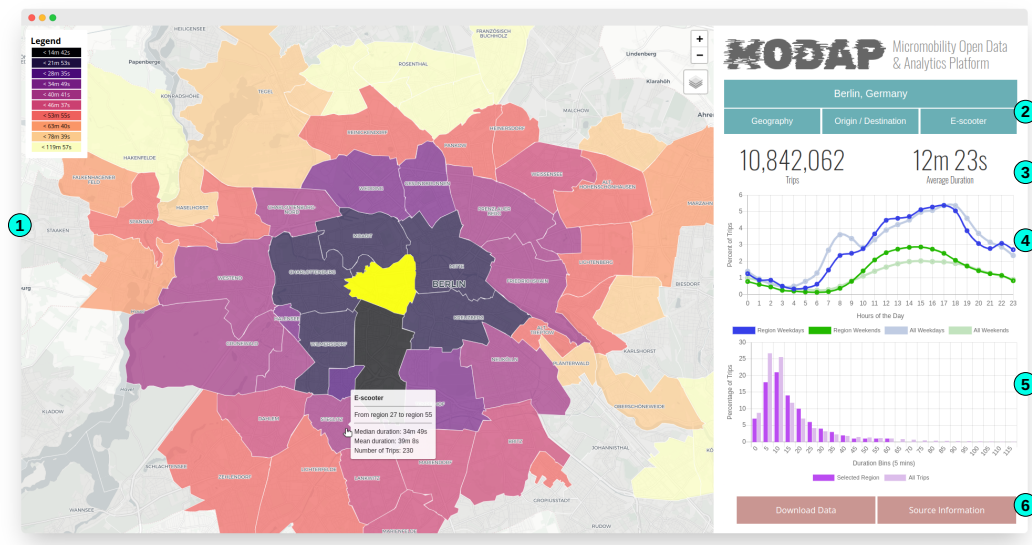


Figure 1 A screenshot of the MODAP web platform located at <https://platial.science/modap>. Features are numbered and described in Section 4.1.

2 Background

Micromobility research has become increasingly important in urban mobility studies as a growing number of cities integrate the services into their transportation ecosystems. At this point, a large body of research has explored various dimensions of micromobility, including regional variability in usage patterns [28], equity implications [7], and the impacts of micromobility on health [2] and safety [29]. This range of topics speaks to the importance of understanding micromobility's role in shaping urban accessibility, sustainability, and public health outcomes.

More broadly, the field of mobility analytics has emerged as a distinct subfield of data analytics [3, 20]. City and transportation planners increasingly rely on sensor-based mobility technologies, and the data they collect, to gain insights into urban activity patterns [9, 25]. Both public transit agencies and ride-hailing companies (e.g., Uber) utilize mobility data and spatial analytics tools to optimize their service delivery and identify where investments and enhancements should be made [23, 11].

Geovisual analytics platforms are a key component of the analytical landscape. Such platforms integrate interactive data visualization, geospatial analysis, and user-focused tools to help stakeholders identify and interpret complex spatiotemporal patterns [5]. These platforms *democratize* data analytics by providing web-based, user-friendly interfaces that require minimal coding expertise or access to specialized computer hardware. A wide range of such platforms are in use today, empowering users to conduct analysis on a variety of topics, from air pollution dynamics [30] or pandemic mobility patterns [6] to social network structures [8] and place-based similarity metrics [18].

The role of *open* data in advancing mobility and urban research is equally important. Open mobility data sets have facilitated transportation research for decades. They have lowered the cost of developing transport services and supported the creation of mobility/transport planning tools. The quintessential example of this is *TriMet*, Portland's transit agency, which in 2005 became one of the first agencies to publish its transit schedule in a machine-readable format. This effort enabled third-party developers (and researcher teams) to build new tools and conduct analyses based on these data [13].

Recently, researchers have put substantial efforts into producing open mobility data sets with the objective of broadening access to critical transportation information [4, 32]. For example, Tenkanen and Toivonen [26] published a longitudinal open travel time data set for multiple mobility modes in Helsinki, Finland, while Kashiya et al. [10] compiled a comprehensive Japan-wide mobility data set from travel surveys. Complementary efforts have focused on establishing data standards and specifications. The Open Mobility Foundation, for instance, works with municipalities, industry stakeholders, and academic researchers to develop standardized mobility data specifications for real-time data sharing.¹

My work builds on these efforts by emphasizing the need for more open mobility data, geovisual exploration, and open analytics platforms. By encouraging more accessible investigation of mobility data, these data and tools contribute not only to advances in geographic information science and transportation planning, but also the broader intersection of mobility, environment, and society.

3 Data & Methodology

3.1 Data collection & cleaning

The data reported through this work were accessed from three dockless micromobility operators, namely *Tier*, *Lime*, and *Flamingo*. Each of these operators runs dockless fleets of vehicles meaning that users can start or end a trip in any public space and the vehicles are not parked at dedicated docking stations. Data from only one operator per city were accessed. Table 1 provides an overview of the final micromobility data sets including the operators, number of trips, type of vehicle(s), and time period of data collection. These trip count values are from after the data has been cleaned.

¹ <https://www.openmobilityfoundation.org/>

■ **Table 1** An overview of the micromobility trip data sets for the five cities.

City	Trip Count	Vehicle Type	Time Period	Operator
Berlin, DE	11,761,219	e-scooter (92%), e-bike	2020-08 – 2024-05	Tier
London, UK	1,192,227	e-scooter (46%), e-bike	2021-08 – 2024-04	Tier
Paris, FR	4,848,310	e-scooter (81%), e-bike	2020-08 – 2024-04	Tier
Washington, D.C., US	5,951,082	e-scooter (67%), e-bike	2022-09 – 2024-05	Lime
Wellington, NZ	1,052,900	e-scooter (100%)	2021-11 – 2024-05	Flamingo

The data were accessed via public-facing application programming interfaces (API). With every request, these APIs (Table 2) return a set of all available vehicles for the requested city, in JSON format. Relevant attribute information include vehicle identifier, geographic coordinates for the current location of the vehicle, vehicle type, and battery level. Each of these APIs were accessed every 60 seconds for the duration of data collection, stated in Table 1. At time of writing, the Lime and Flamingo APIs are still operational, however, Tier merged with another micromobility operator in mid-2024 and discontinued their API at the end of 2024.

■ **Table 2** URLs for the public application programming interfaces associated with each micromobility operator and city.

City	Operator	URL
Berlin, DE	Tier	https://platform.tier-services.io/v2/vehicle?zoneId=berlin
London, UK	Tier	https://platform.tier-services.io/v2/vehicle?zoneId=london
Paris, FR	Tier	https://platform.tier-services.io/v2/vehicle?zoneId=paris
Washington, D.C., US	Lime	https://data.lime.bike/api/partners/v1/gbfs/washington_dc/free_bike_status
Wellington, NZ	Flamingo	https://api.flamingoscooters.com/gbfs/wellington/free_bike_status.json

Provided the set of available vehicles every 60 seconds, trips were identified by noting when a vehicle disappeared from the set of available vehicles (trip start) and when it reappeared in the set of available vehicles (trip end). Given the frequency of requests, this means that trips are accurate to a highest temporal resolution of one minute.

Provided an initial set of trips for each city, the data were then cleaned. Specifically, all trips where the battery level increased between the start and end of a trip were removed from analysis as an increase in battery level suggested that these were recharging/rebalancing trips completed by the operator. Similarly, trips where the average velocity exceeded 20km/hour were removed. All operators in the dataset limit their vehicles to a maximum of 20km/hour. Velocity was calculated as Euclidean distance between origin and destination divided by trip duration. Since full trajectories are not available, average speed is likely underestimated. In addition, trips shorter than 200m or five minutes were removed as well as those longer than 20 km or two hours in duration. This cleaning was done to remove vehicle adjustments and outliers in the data (see [15] for further details).

It is important to mention here that since the emergence of shared micromobility services, the ways in which data have been published via API has changed significantly. Today, many operators obfuscated their vehicle identifiers by randomizing them with every API call. Importantly, for all operators in the data set, I can confirm that the vehicle identifiers are not obfuscated. Lime does obfuscate the identifier in the *vehicle_id* field, but does not for the vehicle identifier in the *rental_uris* parameter. This allows for tracking a vehicle over API requests.

3.2 Data aggregation

Completing the process above resulted in a set of micromobility trip for each of the five cities. Trips included the geographic coordinates of the origin and destination as well as the start time and end time, to the nearest minute.

I then identified three different geospatial units for aggregating the trip data. My motivation for aggregation is presented in the discussion section. The geographies include: 1. Socio-political boundaries. These are sub-city level administration units such as neighborhoods, districts, or traffic analysis zones within a city. As these are determined by the country or city, I understand them to be organized based on the characteristics of the population or physiographic features. 2. Hexagon grid at a 1,000 meter resolution and 3. Hexagon grid at a 500 meter resolution. These two hexagonal grids are uniform geometries that ignore population and physical geography. I felt it important to aggregate at a range of resolutions as these data can be used for different purposes by different stakeholders. The origins and destinations of all trips were intersected with the three different geographies to produce the spatial data sets available for download and analysis.

Temporally, all trips that started or ended within a geographic region were split into either *weekday* or *weekend* and origins were aggregated to the nearest hour. As with the spatial aggregation, this allows for detailed temporal trend analysis but does not allow for the identification of individual trips or users within the data.

■ **Table 3** Description of the contents of *Trip_OD.csv*.

Column	Data Type	Description
gid_o	String	Geographic identifier for the origin of a trip
gid_d	String	Geographic identifier for the destination of a trip
v_type	String	Type of vehicle: scooter or ebicycle
td_mean	Float	Mean duration of trips between two geographies
td_median	Float	Median duration of trips between two geographies
t_count	Integer	Total count of trips between two geographies

Finally, these data were cleaned to remove all geometries that contained no trips and all relevant data were compressed into a series of zipped folders for download. Each zipped directory includes four files: 1. A GeoJSON file containing polygons for the selected geography, 2. A meta data file containing relevant details on the source of the data and provenance information, 3. A *Trip_OD.csv* file containing counts and average duration between pairs of origins and destinations, and 4. A *Region_Details.csv* file containing temporal distribution of trips per geographic region. The data dictionaries for these last two files are presented in Table 3 and Table 4, respectively. These data have been prepared for all five cities and all three geographies and are published under a Creative Commons (CC BY 4.0) license.²

4 Platform

I developed an online, browser-based platform for visual exploration and analysis of the micromobility data. The current platform was designed for a standard computer screens and has not yet been adapted for small screened mobile device. It is available at <https://platial.science/modap> and was designed for two purposes.

² <https://creativecommons.org/licenses/by/4.0/>

■ **Table 4** Description of the contents of *Region_Details.csv*.

Column	Data Type	Description
gid	String	Geographic Identifier
od	String	Origin or destination: “o” or “d”
v_type	String	Type of vehicle: scooter or ebicycle
week	Integer	Weekday (1) or weekend (0)
hour	Integer	Hour of the day (0-23)
td_mean	Float	Mean duration of trips to/from geography
td_median	Float	Median duration of trips to/from geography
t_count	Integer	Total count of trips between to/from geography

The first purpose is to provide an accessible platform for data exploration and pattern discovery. Given the complexity and sheer volume of data, non-technical stakeholders and the general public may find it challenging to navigate. This tool was designed to empower users by allowing them to explore the data independently, compare temporal variations in specific regions against citywide trends, and focus on areas of particular interest.

The second purpose is to provide a platform through which researchers, government agencies, and industry professionals can download the raw data to be use for detailed analyses. Given the volume and different dimensions of the data, I felt it important offer users the ability to first interact with the data through the platform in order to see the mobility data “in-action” before downloading the individual data sets for analysis.

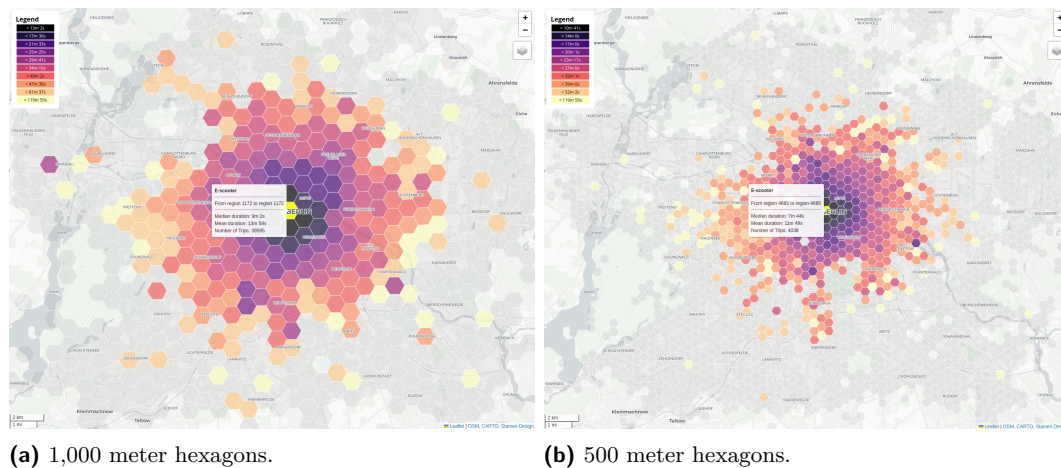
4.1 Features

The MODAP platform boasts a number of features. These features have been numerically labeled in Figure 1 and are referenced in the following descriptive paragraphs in parentheses.

Map Pane (1). The map makes up the majority of the view and is the main point of interaction with the spatial data. This map includes standard tools for panning and zooming, and the ability to change the base map and toggle labels. Once a city and geography are selected, a user is invited to click a geographic region on the map. Once selected, the mobility data are symbolized on the map and a legend is provided. A user can click on the legend to rotate through different color palettes. A user can then hover their mouse over different regions to view details including number and average duration of trips that originate in the selected region (highlighted in yellow) and finishes in the mouse-hovered region.

Selectors (2). On the right side of the platform, there are a set of drop down selectors at the top of the page. These selectors (Green) are comprised of 1. The *City Selector* which allows users to pick a city and is the first point of interaction for the platform. 2. The *Geographies Selector*, allowing users to select from three possible options of geographic polygons (Figure 2). This defaults to administrative regions. 3. The *OD Selector* asks users if they would like to view trips by their origin or by their destination (defaults to origin). 4. When there is more than one vehicle type, the *Vehicle Selector* asks users to select either e-scooters or e-bikes.

General Details (3). Under the Selectors are two large information panes that present the total number of trips per vehicle type for the selected city as well as the average duration for the selected vehicle type in the selected city.



■ **Figure 2** Two of the three geographic units available for spatial analysis. The third is shown in Figure 1.

Hourly Volume Graph (4). A dynamic graph displays trip data based on hour of an aggregated day. The average number of trips (as a percentage of all trips) is shown for weekdays and weekends for (a) the city as a whole and (b) for the geographic region selected on the map. This view permits users to compare selected regions to the overall hourly mobility patterns for the city overall. Users can hover their mouse over points on the graph to view more detailed numbers on the percentage of trips taken per hour.

Duration Graph (5). A dynamic graph showing a histogram of trip durations at five minute intervals. As with the Hourly Volume Graph, the percentage of city trips are shown in a lighter color with the selected geographic region trips shown in a darker purple. This allows users to compare their selected geographic region to the overall city pattern.

Buttons (6). At the bottom right of the screen there are two buttons. The *Data Download* allows users to download the data for the city they selected, aggregated to the geography through which they are viewing the data. The *Source Information* provides metadata related to the data currently being visualized. This includes trip count information, operator, dates, etc.

4.2 Architecture

MODAP is built using completely open source software and the source code is also published as open source at <https://github.com/grantdmckenzie/modap>. On the server, the platform is running a LAPP³ architecture. Specifically, Ubuntu Linux running Apache 2 as the web server. All data are stored in a PostgreSQL relational database with the PostGIS extension allowing for spatial queries. Requests from the client (browser) are handled by a set of PHP scripts. Data downloads have been pre-processed, stored, and shared on the Open Science Framework at <https://github.com/grantdmckenzie/modap>. The front-end is a combination of HTML5, CSS3, and JavaScript. Two JavaScript frameworks are employed for various features. These are *Leaflet*⁴ for the web mapping functionality and *Chart.js*⁵ for the dynamic graphs. All other functionalities were developed with native JavaScript.

³ Linux, Apache, PostgreSQL, PHP

⁴ <https://leafletjs.com>

⁵ <https://chartjs.org>

5 Data showcase

In this section I highlight a few examples of the types of analysis that can be done either through the MODAP web platform or with the open data after downloading.

5.1 Platform-based exploratory analysis

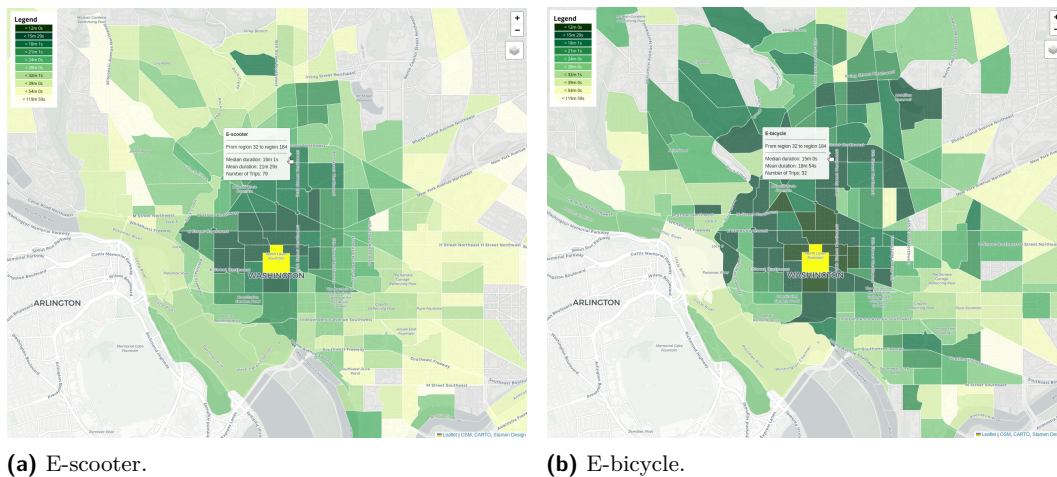


Figure 3 Comparing e-scooter trip duration to e-bicycle trip duration from a selected origin region in Washington, D.C., USA.

As previously mentioned, the platform is designed for visual exploration of micromobility data, aiming to inspire new ideas and facilitate quick analysis of regional and temporal differences across cities, sub-regions, and vehicle types. For instance, one can select a region in a city such as Washington, D.C. and toggle between e-scooters and e-bicycles to identify the difference in trip duration between the selected region and all surrounding regions. Figure 3 shows such a selection. Since the legend remains constant between vehicle types, one can quite easily see that trips taken on e-bicycles travel to further destinations than e-scooters, while maintaining a similar duration. This is especially true for regions in the North of the city. By hovering over different destination regions, one can also view the mean and median duration differences between the two vehicle types.

Other analysis might focus on the difference in micromobility temporal patterns depending on if a region is a trip *origin* or a *destination*. Figure 4 shows a selected region in Wellington, New Zealand. Figure 4a shows the temporal patterns for the region when it is the origin of a trip. The dominant time period for trips originating in this region is during the morning commute on weekdays. We can see that it is higher than the city average, as shown in lighter blue. Compare that to the same region as a destination (Figure 4b). In this case, we find that the dominant time period is evening commuting hours, peaking at 17:00. Based on this exploratory analysis, one might make the initial assumption that the region is zoned to be residential. This might then spark further investigation into land use, socio-economic data, elevation, etc. in the selected and neighboring regions.

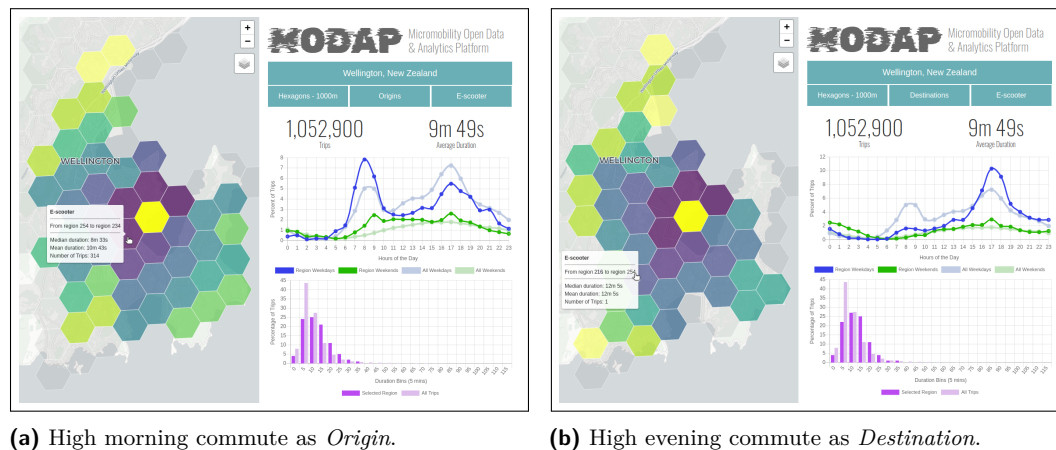


Figure 4 Comparing the difference in origin and destination temporal patterns for a single region in Wellington, New Zealand.

5.2 Analysis with downloaded open data

While the MODAP web platform was designed for visual exploration and analysis of the data, a user also has the ability to download the data sets and run their own analysis offline. These data support a wide range of analytical approaches and research questions. Here, I highlight a few simple examples to illustrate their potential.

One could compare the durations and times across all cities in the data sets to identify the most similar cities and most different cities. For instance, comparing various e-vehicle behavior has been a topic of interest recently [1, 17]. Through these open data, one can identify the difference in trip durations between e-scooters and e-bicycles in London, United Kingdom, for example. Figure 5 shows a density plot of each vehicle's trip durations. This could be compared to other cities and used to inform policymakers on the suitability of these different vehicle types in their cities.

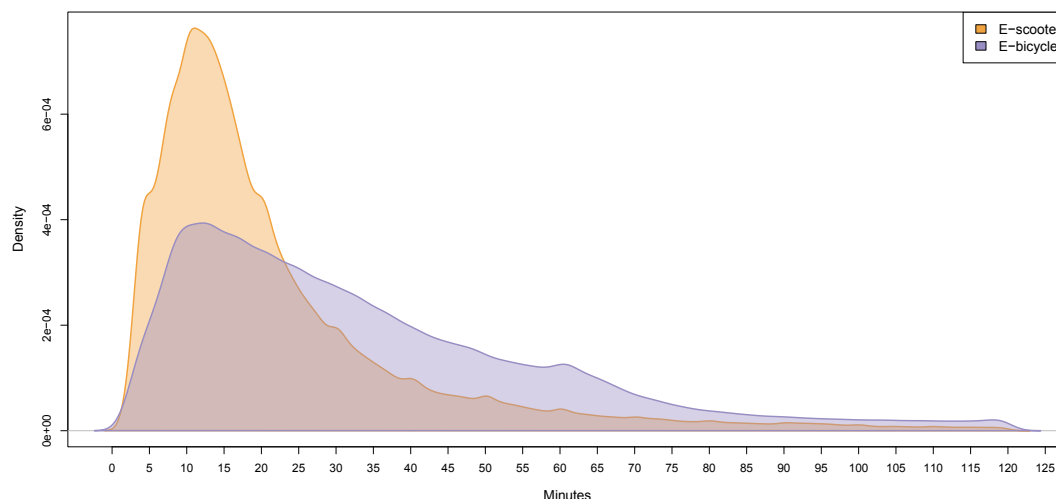


Figure 5 Density plots comparing e-scooter and e-bicycle trip duration in London, UK.

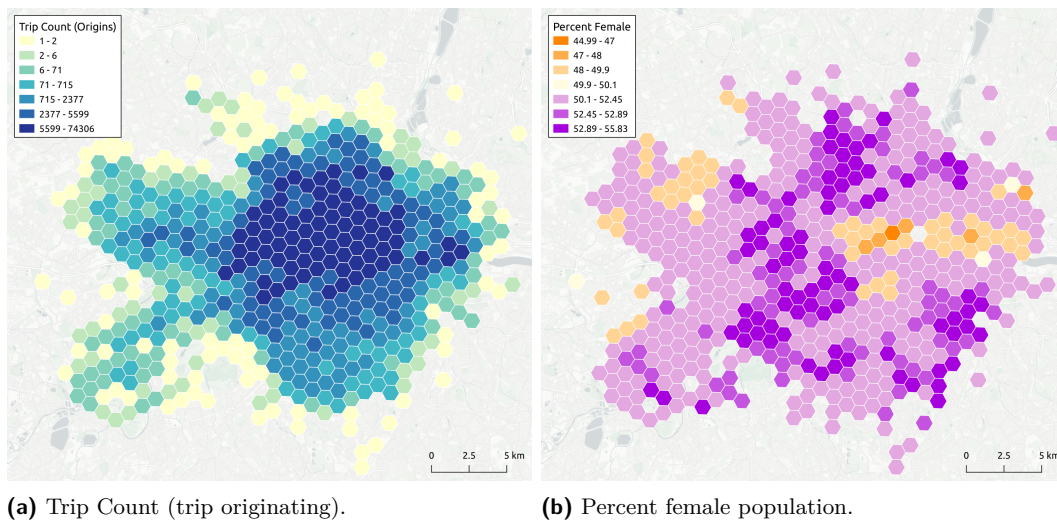


Figure 6 Comparison between micromobility trip counts and percent female population in London, aggregated to 1000 m hexagon cells.

One could also use these data to identify the relationship between sex (male and female)⁶ and micromobility trip volume in London, United Kingdom (Figure 6). To accomplish this, I accessed the demographic data from the UK 2021 Census⁷ at the Middle layer Super Output Areas (MSOA) level and ran an areal interpolation [24] of the MSOA's data to assign male and female population counts to the same 1,000 meter hexagon geography published through MODAP. Then, having data at the same geospatial resolution, I ran a Bivariate Moran's *I* analysis comparing trip volume and percent female population in the city of London. The results report a Bivariate Moran's *I* value of -0.2043 ($p < 0.05$) with a z-score of -5.8137. This suggests that there is a significant inverse spatial relationship between the two data sets, or rather that there is a strong relationship between micromobility trips and regions with higher male residential populations. This simple analysis is meant to demonstrate how these data could be used and further study might investigate additional factors related to the built environment and other socio-economic factors.

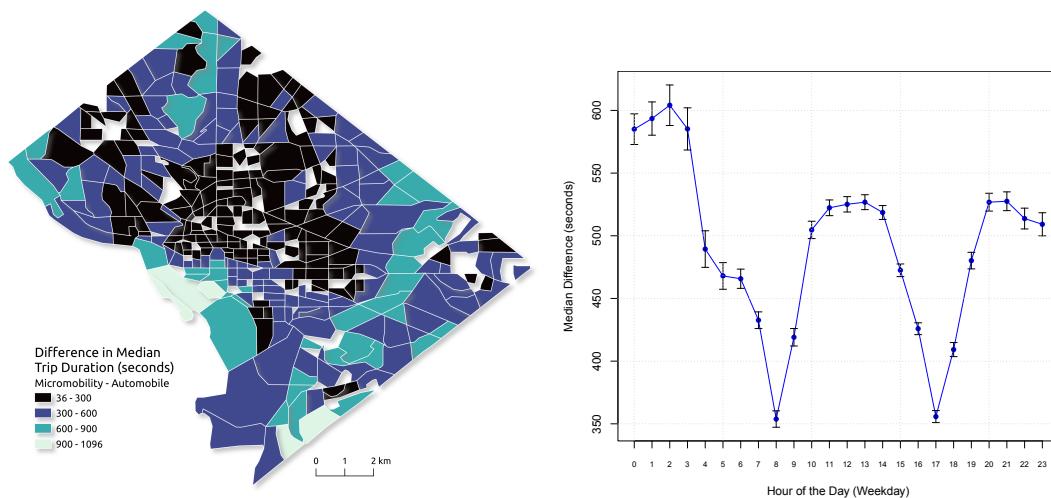
Finally, one could compare micromobility usage to automobile usage in a city. In this example I investigate spatial and temporal differences in duration of trips taken by the two modes of travel. The ride hailing company, Uber, previously published trip duration data at the level of Traffic Analysis Zones (TAZ) for a number of global cities through their Uber Movement platform.⁸ With access to these data, we can compare micromobility trip durations to those of passenger automobile trips. The Uber Movement data reports mean and median trip duration between all possible TAZ at every hour of either a weekday or weekend. Given the data in MODAP is published at the same spatial and temporal resolution (Administrative TAZ and hourly) for Washington, D.C., we can compare the two modes of travel. Figure 7a shows the difference in median trip duration between the two modes of travel for each origin TAZ in the city. This is calculated by subtracting the median automobile duration from the median micromobility duration for each TAZ. The results are

⁶ Reported as sex identified at birth in the UK Census.

⁷ <https://www.ons.gov.uk/census>

⁸ <https://www.uber.com/en-CA/blog/kepler-data-visualization-traffic-safety/>

the difference reported in seconds. We can see that there is some spatial clustering with the downtown core reporting the lowest difference and those on the outskirts of the city reporting larger differences. This is in line with existing work [16] comparing these data five years prior. Figure 7b reports these differences in duration by hour of a standard weekday. In this case, median automobile trip duration between all pairs of TAZ are subtracted from median micromobility trip duration and then the median of all of those is reported for each hour. We can see that while automobile trips are faster in all cases (y-axis is difference in seconds), the two modes of transport become more similar during peak commuting hours. This analysis showcases the types of comparison analysis that can be conducted between micromobility and other modes of transport within cities.



(a) Difference in average trip duration for each traffic analysis zone. Median micromobility trip duration minus median automobile trip duration.

(b) Median difference in trip duration by hour of the weekday. Median of micromobility trip duration minus automobile trip duration. Error bars show standard error.

Figure 7 Spatial and temporal comparison of micromobility trip duration to automobile trip duration in Washington, D.C., USA.

6 Discussion & conclusions

The main objective of this paper is to highlight the importance of open access to high quality mobility data as well as the exploratory tools necessary to analyze them. While this work is narrowly focused on the origins, destinations, and durations of micromobility trips in five major cities, it adds to the growing availability of open data that urban planners, academic researchers, and the public can use to better understand mobility around the world. It is my intention to continue to contribute to this data set and platform as I collect and curate additional e-scooter, e-bicycle, and other sources of micromobility data (e.g., e-moped, non-e-bicycle).

One question that I feel it was important to address is, why I did not choose to publish the *raw* trip data. It took a significant amount of time and effort to collect these data over the course of multiple years and during that time I realized that there are aspects of the trip data that are incredibly sensitive from an individual privacy perspective. Unlike public transit or docking station-based micromobility, users of dockless micromobility vehicles

typically park their vehicles directly outside of their homes or near places of interest that host activities that expose personal information about their clientele. I am also aware of numerous legal cases related to the personal privacy of individual micromobility trip data [12]. In these cases, either precise geographic coordinates of origins and destinations are shared or raw trajectories. As researchers with access to these highly sensitive data, I felt an ethical responsibility not to publish the raw data, but instead preserve a level of privacy through spatial and temporal aggregation [19].

As researchers, I also felt it was important to publish a public data set for the variety of stakeholders who could significantly benefit from access to mobility patterns. When working with data such as these there is always a trade-off between privacy and utility [31]. The size and variety of the geographies chosen, the aggregation to hours of a day were all selected with individual privacy in mind. My approach to aggregating the data is far from perfect, but it does protect privacy while still offering valuable insight from the mobility patterns. My team and I will continue to explore alternative geoprivacy preservation techniques with the objective of releasing higher spatial and temporal resolution data.

There are a number of areas for improvement in both the platform and the data. First, each city is only represented by a single micromobility operator and for most of the cities selected for this project, there are multiple operators. While minimal, existing research has demonstrated that there are differences between operators within the same city [16]. Second, my analysis includes three operators over five cities making it difficult to compare both cities and operators at the same time. Finally, I have no details on the actual riders themselves, nor the purpose for their trips. While this is arguably outside the scope of this specific project, not having access to user demographics or trip purpose limits the forms of analysis that can be done with these data.

My future work on this project will involve expanding the open data set to numerous other cities and micromobility operators. My intention is also to allow users to upload their own data sets to the MODAP platform in order to visually and statistically compare patterns and to help users clean their own mobility data.

Conclusion

Micromobility services have grown substantially in recent years, now constituting a not-insignificant share of short, urban trips. As these services expand, they contribute to the evolving landscape of urban mobility, offering an alternative to, and often complementing, traditional transportation modes. In most cities, there is a regulatory requirement that operators of these services provide open APIs that publish real-time data on vehicle availability. In this work, I leveraged these data to reconstruct trips across five major cities over a three-year period (in most cases). Through the MODAP project, I am making these trip data openly available for download and analysis. To support these data, I developed an interactive geovisualization platform that enables users to engage and explore these data through their web browser. My objective is to provide researchers, policymakers, and urban planners with a rich and open source of new mobility data. My intention is for this to support evidence-based decision-making and contribute to the broader discourse on sustainable and equitable urban mobility.

References

- 1 Mohammed Hamad Almannaa, Huthaifa I Ashqar, Mohammed Elhenawy, Mahmoud Masoud, Andry Rakotonirainy, and Hesham Rakha. A comparative analysis of e-scooter and e-bike usage patterns: Findings from the city of Austin, tx. *International Journal of Sustainable Transportation*, 15(7):571–579, 2021.
- 2 Alexandra Bretones and Oriol Marquet. Riding to health: Investigating the relationship between micromobility use and objective physical activity in Barcelona adults. *Journal of Transport & Health*, 29:101588, 2023.
- 3 Vanessa Brum-Bastos and Antonio Páez. Hägerstrand meets big data: time-geography in the age of mobility analytics. *Journal of Geographical Systems*, 25(3):327–336, 2023. doi:10.1007/S10109-023-00421-0.
- 4 Mollie C D’Agostino, Paige Pellaton, and Austin Brown. Mobility data sharing: challenges and policy recommendations. Technical report, UC Davis: Institute of Transportation Studie, 2019.
- 5 Leonardo Ferreira, Gustavo Moreira, Maryam Hosseini, Marcos Lage, Nivan Ferreira, and Fabio Miranda. Assessing the landscape of toolkits, frameworks, and authoring tools for urban visual analytics systems. *Computers & Graphics*, 123:104013, 2024. doi:10.1016/J.CAG.2024.104013.
- 6 Song Gao, Jinmeng Rao, Yuhao Kang, Yunlei Liang, and Jake Kruse. Mapping county-level mobility pattern changes in the United States in response to covid-19. *SIGSpatial Special*, 12(1):16–26, 2020. doi:10.1145/3404111.3404115.
- 7 Xiaodong Guan, Dea van Lierop, Zihao An, Eva Heinen, and Dick Ettema. Shared micromobility and transport equity: A case study of three european countries. *Cities*, 153:105298, 2024.
- 8 Sichen Jin, Alex Endert, and Clio Andris. Snoman: a visual analytic tool for spatial social network mapping and analysis. *Cartography and Geographic Information Science*, pages 1–19, 2024.
- 9 Jens Kandt and Michael Batty. Smart cities, big data and urban policy: Towards urban analytics for the long run. *Cities*, 109:102992, 2021.
- 10 Takehiro Kashiyama, Yanbo Pang, Yuya Shibuya, Takahiro Yabe, and Yoshihide Sekimoto. Nationwide synthetic human mobility dataset construction from limited travel surveys and open data. *Computer-Aided Civil and Infrastructure Engineering*, 2024.
- 11 Zhitao Li, Jinjun Tang, Tao Feng, Biao Liu, Junqiang Cao, Tianjian Yu, and Yifeng Ji. Investigating urban mobility through multi-source public transportation data: A multiplex network perspective. *Applied Geography*, 169:103337, 2024.
- 12 Jennifer Lynch. EFF, ACLU Urge Appeals Court to Revive Challenge to Los Angeles’ Collection of Scooter Location Data, 2021. URL: <https://www.eff.org/press/releases/eff-aclu-urge-appeals-court-revive-challenge-los-angeles-collection-scooter-riders>.
- 13 Bibiana McHugh. Pioneering open data standards: The gtfs story. *Beyond transparency: open data and the future of civic innovation*, pages 125–135, 2013.
- 14 Grant McKenzie. grantdmckenzie/modap. Software, swhId: swh:1:dir:a48742b027327af1a11426dcff7ddf77efa10fb9 (visited on 2025-08-04). URL: <https://github.com/grantdmckenzie/modap>, doi:10.4230/artifacts.24224.
- 15 Grant McKenzie. Spatiotemporal comparative analysis of scooter-share and bike-share usage patterns in washington, dc. *Journal of transport geography*, 78:19–28, 2019.
- 16 Grant McKenzie. Urban mobility in the sharing economy: A spatiotemporal comparison of shared mobility services. *Computers, Environment and Urban Systems*, 79:101418, 2020. doi:10.1016/J.COMPENVURBSYS.2019.101418.
- 17 Grant McKenzie. A comparison of electric and non-electric bike sharing in montreal, canada. In *Proceedings of the 18th International Conference on Computational Urban Planning and Urban Management*, 2023. doi:10.17605/USF.IO/6YR5V.

- 18 Grant McKenzie, Sarah Battersby, and Vidya Setlur. Mixmap: A user-driven approach to place-based semantic similarity. *Cartography and Geographic Information Science*, 51(4):583–598, 2024.
- 19 Grant McKenzie, Hongyu Zhang, and Sébastien Gambs. Privacy and ethics in geoai. In *Handbook of Geospatial Artificial Intelligence*, pages 388–405. CRC Press, 2023.
- 20 Harvey J Miller. Movement analytics for sustainable mobility. *Journal of Spatial Information Science*, 1(20):115–123, 2020. doi:10.5311/JOSIS.2019.20.663.
- 21 North American Bikeshare and Scootershare Association. 2023 shared micromobility state of the industry report. <https://nabsa.net/industry/>, 2024. Accessed: 2025-01-31.
- 22 John Pflueger. Driving positive outcomes through open data solutions for mobility | dell. Technical report, Dell, 2018.
- 23 Divya Babu Ravichandran and Varun Verma. How Data Shapes the Uber Rider App. Technical report, Uber, 2021. URL: <https://www.uber.com/en-CA/blog/how-data-shapes-the-uber-rider-app/>.
- 24 Sergio J Rey and Luc Anselin. Pysal: A python library of spatial analytical methods. In *Handbook of applied spatial analysis: Software tools, methods and applications*, pages 175–193. Springer, 2009.
- 25 Nicolas Tempelmeier, Yannick Rietz, Iryna Lishchuk, Tina Kruegel, Olaf Mumm, Vanessa Miriam Carlow, Stefan Dietze, and Elena Demidova. Data4urbanmobility: Towards holistic data analytics for mobility applications in urban regions. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 137–145, 2019. doi:10.1145/3308560.3317055.
- 26 Henrikki Tenkanen and Tuuli Toivonen. Longitudinal spatial dataset on travel times and distances by different travel modes in helsinki region. *Scientific data*, 7(1):77, 2020.
- 27 United Nations - Population Division. World urbanization prospects 2018. <https://www.un.org/development/desa/pd/news/world-urbanization-prospects-2018>, 2018. [Accessed 31-01-2025].
- 28 Priyanka Verma and Grant McKenzie. Regional comparison of socio-demographic variation in urban e-scooter usage. *Environment and Planning B: Urban Analytics and City Science*, 51(7):1548–1562, 2024.
- 29 Hong Yang, Qingyu Ma, Zhenyu Wang, Qing Cai, Kun Xie, and Di Yang. Safety of micro-mobility: Analysis of e-scooter crashes by mining news reports. *Accident Analysis & Prevention*, 143:105608, 2020.
- 30 Xiaoqi Yue, Dan Feng, Desheng Sun, Chao Liu, Hongxing Qin, and Haibo Hu. Airpollutionviz: visual analytics for understanding the spatio-temporal evolution of air pollution. *Journal of Visualization*, 27(2):215–233, 2024. doi:10.1007/S12650-024-00958-2.
- 31 Hongyu Zhang and Grant McKenzie. Rehumanize geoprivacy: From disclosure control to human perception. *GeoJournal*, 88(1):189–208, 2023.
- 32 Kai Zhao, Sasu Tarkoma, Siyuan Liu, and Huy Vo. Urban human mobility data mining: An overview. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 1911–1920. IEEE, 2016. doi:10.1109/BIGDATA.2016.7840811.

A Modularity-Driven Framework for Unraveling Congestion Centers with Enhanced Spatial-Semantic Features

Weihua Huan ✉ 

College of Surveying and Geo-informatics, Tongji University, Shanghai, China

Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Kowloon, Hong Kong

Xintao Liu ✉ 

Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Kowloon, Hong Kong

Wei Huang¹ ✉  

College of Surveying and Geo-informatics, Tongji University, Shanghai, China

Department of Civil Engineering, Toronto Metropolitan University, Canada

Urban Mobility Institute, Tongji University, Shanghai, China

Abstract

The propagation of traffic congestion is a complicated spatiotemporal phenomenon in urban networks. Extensive studies mainly relied on dynamic Bayesian network or deep learning approaches. However, they often struggle to adapt seamlessly to diverse data granularities, limiting their applicability. In this study, we propose a modularity-driven method to unravel the spatiotemporal congestion propagation centers, effectively addressing temporal granularity challenges through the use of the fast Fourier Transform (FFT). Our framework distinguishes itself due to its capacity to integrate enhanced spatial-semantic features while eliminating temporal granularity dependence, which consists of two data-driven modules. One is adaptive adjacency matrix learning module, which captures the spatiotemporal relationship from evolving congestion graphs by fusing node degree, spatial proximity, and the FFT of traffic state indices. The other one is local search module, which employs local dominance principles to unravel the congestion propagation centers. We validate our proposed methodology on the large-scale traffic networks in New York City, the United States. An ablation study on the dataset reveals that the combination of the three features achieves the highest modularity scores of 0.65. The contribution of our work is to provide a novel way to infer the propagation centers of traffic congestion, and reveals the flexibility of extending our framework at temporal scales. The network resilience and dynamic evolution of the identified congestion centers can provide implications for actional decisions.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases Congestion center, Temporal granularity, Fast Fourier Transform, Local dominance

Digital Object Identifier 10.4230/LIPIcs.GIScience.2025.7

Supplementary Material *Software*: <https://figshare.com/s/55638fbbd7f4c59a419c>

Acknowledgements I want to thank my supervisor Wei Huang from Tongji University, China and co-supervisor Xintao Liu from the Hong Kong Polytechnic University, China.

¹ Wei Huang is the corresponding author



© Weihua Huan, Xintao Liu, and Wei Huang;
licensed under Creative Commons License CC-BY 4.0

13th International Conference on Geographic Information Science (GIScience 2025).

Editors: Katarzyna Sila-Nowicka, Antoni Moore, David O'Sullivan, Benjamin Adams, and Mark Gahegan;
Article No. 7; pp. 7:1–7:11



Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

Traffic congestion is a pervasive issue in urban road networks, propagating across both time and space. With the exponential growth of multi-source real-time geographic data enriched with temporal and spatial information, significant efforts have been made to uncover the spatiotemporal patterns of traffic congestion propagation. Existing studies can be broadly categorized into two main approaches: dynamic Bayesian network (DBN) [14] and deep learning models [1]. DBN-based approaches discretize continuous historical traffic data into discrete traffic states and then infer congestion propagation patterns by calculating the state transition probability between adjacent road segments [4, 5, 11]. However, these approaches are often limited by the loss of information during data discretization and the reliance on prior knowledge, which may compromise the accuracy of congestion propagation inference. On the other hand, deep learning models, such as graph neural network (GNN) [17] and graph convolutional network (GCN) [6, 21], leverage road network topology to construct feature matrices for congestion analysis. While these methods have shown promise, they predominantly rely on predefined adjacency matrices based on simplistic metrics such as node connectivity or spatial proximity. This static representation fails to capture the dynamic and adaptive nature of congestion propagation, which evolves over time and is influenced by multiple factors beyond mere spatial relationships.

To address these limitations, the objective of this study is to propose a modularity-driven framework to track the evolution of traffic congestion propagation centers with enhanced node feature fusion. Our approach introduces two key innovations that significantly advance the state-of-the-art in congestion center analysis.

- Adaptive multi-feature fusion adjacency matrix. Unlike traditional methods that rely on static adjacency matrices, we design an adaptive multi-feature adjacency matrix that integrates enhanced node features - including degree, spatial proximity, and the fast Fourier Transform (TTF) of the traffic state index (TSI) – to capture the complex interplay of factors driving congestion propagation. This matrix dynamically updates over time, enabling a more accurate representation of the information-passing process between road segments at different timestamps.
- Fast Fourier Transform of the TSI to eliminate temporal-scale effects: To address the potential impact of varying temporal resolutions (e.g., 5 minutes, 30 minutes, or 1 hour) on model performance, we introduce the FFT of the TSI. This transformation eliminates the influence of time scales, ensuring that our model remains robust and effective across different data granularities. This innovation is particularly critical for real-world applications where data collection intervals may vary.

The significance of our work lies in its ability to provide a flexible framework for identifying congestion propagation centers, which is independent of temporal granularity thus can be extended at multiple scales. Experimental results on the traffic floating car datasets from New York City (NYC), the United States, demonstrate the effectiveness of our method based on the enhanced features, achieving an average modularity score level at 0.65. Besides, the propagation probability and distribution of the congestion centers on different types of days (i.e., weekdays, weekends, and holidays) are further visualized. These findings highlight the potential of our framework to advance the traffic congestion propagation analysis and provide actionable insights for urban transportation management.

The remainder of this paper is organized as follows. **Section 2** reviews the related works and identifies some limitations. **Section 3** provides the framework and details of our proposed methodology. **Section 4** presents the experiment and results. **Section 5** concludes this paper.

2 Related works

Significant efforts have been devoted to traffic congestion propagation modeling, with existing approaches primarily falling into two categories: dynamic Bayesian network (DBN) and deep learning. On the one hand, DBN-based methods model traffic congestion propagation as state transitions between adjacent road segments in a probabilistic graph [14]. For instance, Nguyen et al. (2016) [11] constructed causal congestion trees using taxi trajectory data to estimate causality probabilities and reveal the interactions of traffic streams. Building upon this, Chen et al. (2018) [4] proposed the spatiotemporal congestion subgraph (STCS) based on travel time data to describe recurring congestion propagation patterns. Similarly, Fan et al. (2019) [5] developed a DBN-based prediction model using floating car data, discretizing traffic speed to predict congestion diffusion states. However, these DBN-based approaches heavily rely on prior knowledge and data discretizations, which may lead to information loss and reduced inference accuracy.

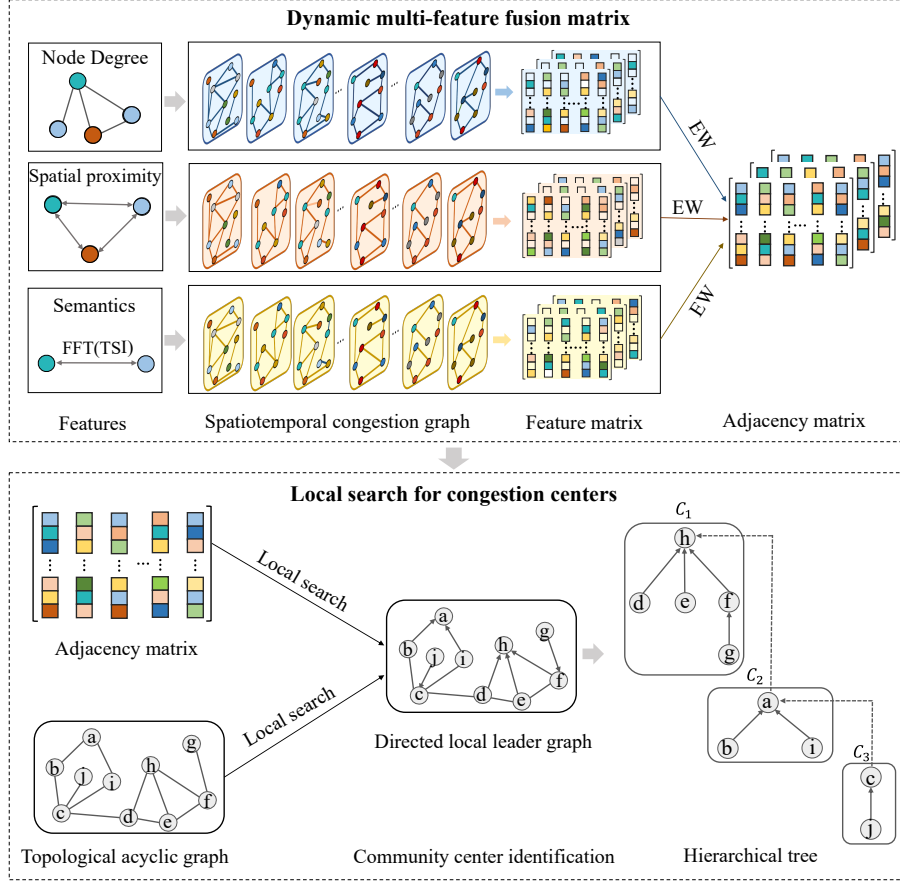
On the other hand, deep learning models treat congestion propagation as a multi-dimensional feature learning problem. Representative traditional models, such as LSTM [2] and LSTM variants [15, 19], have been widely applied. However, these models often fail to incorporate road network topology, resulting in suboptimal performance. To address this limitation, graph-based approaches, particularly graph convolutional networks (GCNs) (Kipf and Welling, 2016), have gained traction. GCNs model traffic networks as graphs, where road segments are represented as nodes, enabling the direct incorporation of spatial dependencies. For example, Zhao et al. (2019) [20] proposed temporal GCN (T-GCN) to capture both spatial and temporal dependencies, while Liang et al. (2022) [8] introduced spatiotemporal GCN (ST-GCN) to simultaneously model spatiotemporal dependencies and heterogeneity. Zheng et al. (2022) [21] further advanced this field by proposing dynamic STGCN (DST-GCN), which constructs spatial-temporal graphs across time slices by connecting the latest time slice with past slices. Notably, Luan et al. (2022) [9] integrated Bayesian inference with deep learning to develop a dynamic Bayesian graph convolutional network (DBGCN). Despite this, temporal granularities of the data potentially affect model effectiveness. In addition to DBNs and deep learning models, some studies have introduced epidemiological models, such as the susceptible-infectious-recovered (SIR) model, to analyze congestion propagation [7, 12, 16]. While these models provide valuable insights, they often rely on simplifying assumptions and do not fully account for the complex topology of urban road networks. Furthermore, the effectiveness of above methods highly relies on the spatiotemporal resolution of collected data, as it seeks to reveal large-scale traffic dynamics by obscuring small-scale spatiotemporal interactions [18].

3 Proposed methodology

3.1 Framework

The framework of the proposed method contains two-driven modules (Fig. 1): dynamic adjacency matrix learning module and local search module for congestion centers detection. In the first module, adaptive adjacency matrices are constructed by integrating three key features: node degree, spatial proximity, and semantic information derived from traffic spatiotemporal congestion graphs (detailed in **Section 3.2**). These matrices dynamically capture the relationships between nodes in the traffic network, reflecting both structural and contextual properties of congestion patterns. The generated adjacency matrices are then input into the second module, where a local search algorithm (**Section 3.3**) is applied to identify

congestion centers at multiple scales (e.g., C_1 , C_2 , and C_3 in Fig. 1). This hierarchical detection process enables the framework to uncover the centers of congestion propagation process across different spatial resolutions, providing a comprehensive understanding of traffic dynamics.



■ **Figure 1** Framework of the proposed method.

3.2 Dynamic multi-feature fusion adjacency matrix

To enhance the discriminative capability of relative closeness in multi-attribute decision-making, the entropy weight (EW) method is employed, as it effectively balances the contribution of diverse attributes. Leveraging this advantage, the EW approach is utilized in this study to compute adaptive adjacency matrices. Given a spatiotemporal congestion subgraph G_{t_j} at timestamp t_j , its adaptive adjacency matrix M_{t_j} is derived through a weighted fusion of its degree similarity D_{t_j} , Spatial proximity similarity S_{t_j} , and the FFT of traffic state similarity F_{t_j} :

$$M_{t_j} = w_{t_j,1}D_{t_j} + w_{t_j,2}S_{t_j} + w_{t_j,3}F_{t_j}, \quad (1)$$

where $w_{t_j,1}$, $w_{t_j,2}$, and $w_{t_j,3}$ denote the EWs calculated by the information entropy of each similarity matrix. The degree similarity D_{t_j} is computed based on the Cosine similarity of node degrees. FFT converts TSI in temporal domain into a frequency domain signal, which is calculated as:

$$F_{t_j}(k) = \sum_{n=0}^{N-1} TSI_{t_j}(n) \bullet e^{-i\frac{2\pi}{N}kn}, \quad k = 1, 2, \dots, N, \quad (2)$$

$$TSI_{t_j} = \frac{v - \bar{v}_{t_j}}{v}, \quad (3)$$

where N is the length of time series TSI_{t_j} . $TSI_{t_j}(n)$ is the TSI value at the n th dimension. $e^{-i\frac{2\pi}{N}kn}$ is the kernel of Fourier Transform, representing the complex exponential signal with frequency $\frac{k}{N}$. v represents the free-flow speed of road segment, \bar{v}_{t_j} is the actual average speed at timestamp t_j . The range of TSI_{t_j} is $[0,1]$, and the threshold is 0.7 [4]. Therefore, the road segment at the timestamp t_j is defined as a spatiotemporal congestion instance if its TSI_{t_j} is no less than 0.7. Given a spatiotemporal congestion graph G , its adjacency matrices across J timestamps are dynamically updated as $M = [M_{t_1}, M_{t_2}, \dots, M_{t_J}]$, ensuring a time-sensitive representation of congestion propagation patterns.

3.3 Local search algorithm

Based upon the adaptive adjacency matrices, a local search algorithm is employed to detect multi-scale communities in dynamic networks. The process of local search in this study is shown as the local search module in Fig. 2, which can be broken into the following four stages: (i) Node value assignment. Each node u is assigned a value x_u by summing the weights of its connected edges, derived from the dynamic adjacency matrix and the spatiotemporal congestion graph. This creates a directed acyclic graph where each node points to its highest-value neighbor, provided that neighbor's value is greater than or equal to its own. (ii) Local leader identification. Local leaders [3] are identified as nodes with incoming edges but no outgoing edges, representing dominant points in the network's community structure (e.g., nodes a, c, and h in Fig. 1). These leaders help reveal the network's hierarchical organization. (iii) Local breath-first search [13]. For each local leader u , an LBFS algorithm is used to find the nearest local leader v with $x_u \leq x_v$. LBFS is efficient, stopping once the nearest leader is found, and provides the shortest path length between leaders, offering insights into network connectivity. (iv) Multi-scale community detection. multi-scale communities (denoted as the symbol C_i , where i represents different levels or scales) are identified, capturing the network's structure at varying resolutions.

4 Case studies

4.1 Datasets

The study area is New York City, the United States. The NYC floating car data was downloaded from Uber Movement, covering a time period from December 1, 2018 to December 31, 2018. The time interval is 1 hour. Each record contains recording time, road segment ID, and average speed. The free-flow speed is also acquired from the Uber Movement. It equals to the 15th percentile value of the actual speeds of all floating vehicles on a road segment, with speeds sorted in descending order. Therefore, a 24-dimensional time series feature over one day can be obtained.

4.2 Experiment results

4.2.1 Evaluation metrics

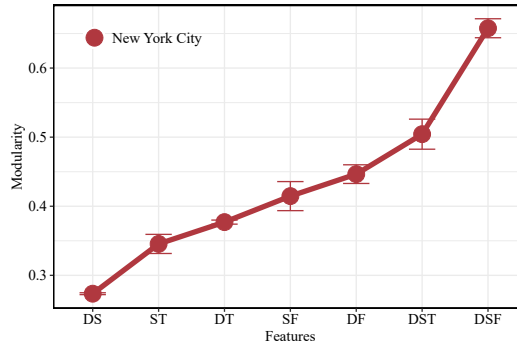
Modularity is a commonly used indicator that measures the quality of community division, which is employed as the evaluation metric in this study. The selection of modularity is appropriate for this study as it is a mature index to effectively quantify community structure without predefined labels - a critical advantage given the lack of verified community partitions in real transportation networks. The modularity, denoted as Q , is defined as the difference between the proportion of intra community edges with the expected number of such edges in a random graph with identical degree sequence [10], formulated as:

$$Q = \frac{1}{2m} \sum_{xy} \left[A_{ij} - \frac{k_i k_j}{2m} \delta(c_i c_j) \right], \quad (4)$$

where m is the total number of edges in the network. A_{ij} represents the weight of the edge between nodes i and j . k_i and k_j are the sum of the weights of the edges attached to nodes i and j . $\delta(C_i, C_j) = 1$ if i and j belong to the same community (i.e., $C_i = C_j$). Otherwise, $\delta(C_i, C_j) = 0$. The range of Q is $[-1/2, 1]$. Q greater than 0.5 means the results are convincing.

4.2.2 Ablation study

An ablation experiment was conducted to evaluate the impacts of different attribute combinations on model performance, i.e., node degree (D), spatial proximity (S), TSI (T), and the fast Fourier transform of the TSI (F). By utilizing the EW method, seven types of adaptive adjacency matrices were calculated (i.e., DS, DT, DF, ST, SF, DST, and DSF). Fig. 2 displays the modularity based on the seven adaptive adjacency matrices: $Q_{DSF} > Q_{DST} > Q_{DF} > Q_{SF} > Q_{DT} > Q_{ST} > Q_{DS}$. Some findings can be concluded: (i) Q_{DSF} secures the highest value, emphasizing the combination of spatial (i.e., node degree and spatial proximity) and semantic information (i.e., Fast Fourier Transform of the TSI) works best. (ii) $Q_{DSF} > Q_{DST}$ proves that the Fast Fourier Transform of the TSI improves the modularity compared to the original TSI. (iii) $Q_{DF} > Q_{SF}$ and $Q_{DT} > Q_{ST}$ suggest that the node degree is more useful than spatial proximity to detecting well-structured communities. This is because node degree captures more inherent structural details in the time series data. (iv) Q_{DS} is below 0.3, indicating that relying on spatial information, without incorporating semantic information, struggles to accurately identify communities. Based upon these results, we can rank the significance of the attributes for our model: $F > T > D > S$. The results provide empirical evidence that effective community detection requires both multi-scale structural analysis and sophisticated semantic information, instead of feature inclusion without discrimination.



■ **Figure 2** Ablation study of the proposed method based on the different feature combinations.

4.2.3 Propagation pathway

In this section, we analyze how the searched congestion communities aligns with human travel patterns throughout the day based on the DSF. To realize this, the communities for each hours are firstly grouped based on the type of days (i.e., weekdays, weekends, and holidays). Then the same communities between continuous time periods (i.e., $t_1 - t_5$) are counted as the number of community transfers. The transfer values are normalized to a ratio p between 0 and 1, visualized as the lines in in Fig. 3.

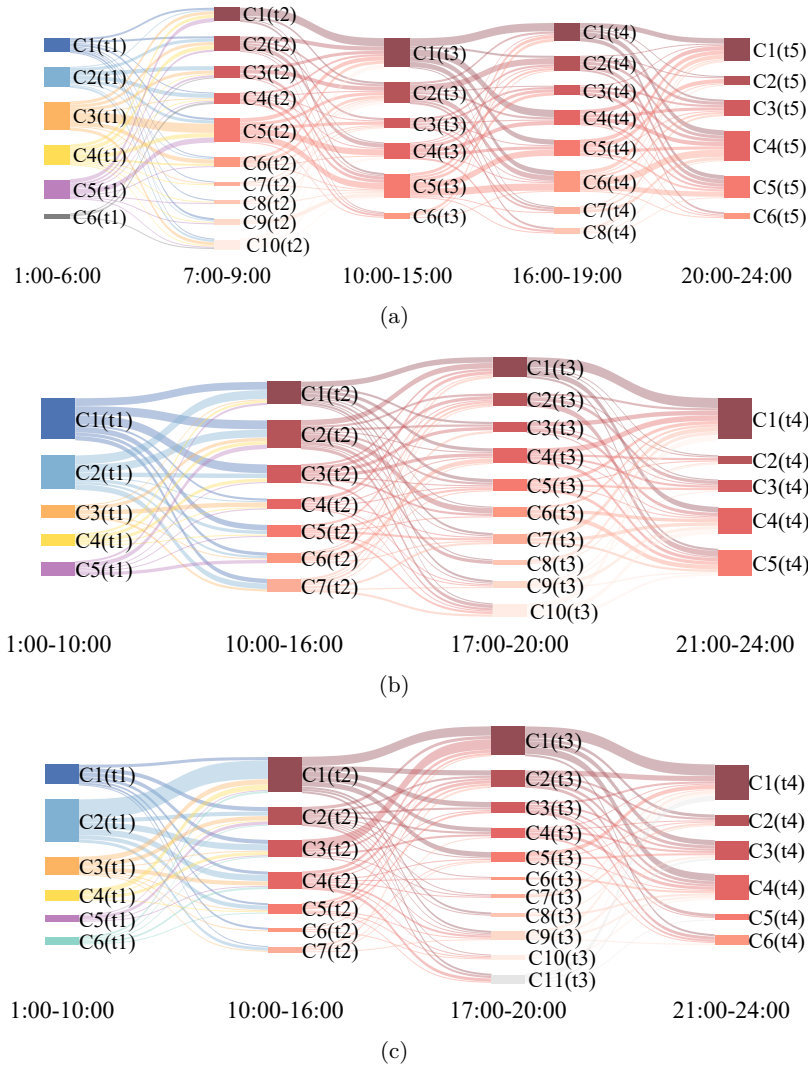


Figure 3 Congestion community transfer during the same-type days.

The nodes represent community centers, labeled as C_i at timestamp t_j ($i \in [1, 11]$, $j \in [1, 5]$). The size of the node C_i shows its proportion p_i , and the thickness of the connecting lines reflects the transfer strength. Some key findings are concluded: The scale of communities at peak-hours on weekdays are growing from congestion bottleneck during 1:00 to 6:00. However, the size of $C1(t_2)$ and $C1(t_4)$ does not evolve very drastically compared with $C1(t_1)$, implying that the increased human travel usually results in an increase in

small-scale communities, but rarely changes the primary communities. This rule also can be observed on weekends and holidays. Besides, by comparing the size of communities at t_1 and t_5 , the relative proportions of communities at different levels are similar. This reflects the “self-regulation” of congestion bottlenecks.

Subsequently, we visually show how the detected congestion centers distribute over time at 71 community districts. Fig. 4 shows the occurrence frequency and distribution of congestion centers at peak hours on weekdays, weekends, and holidays, which are captured from dynamic evolving videos of the community centers at specific time periods.

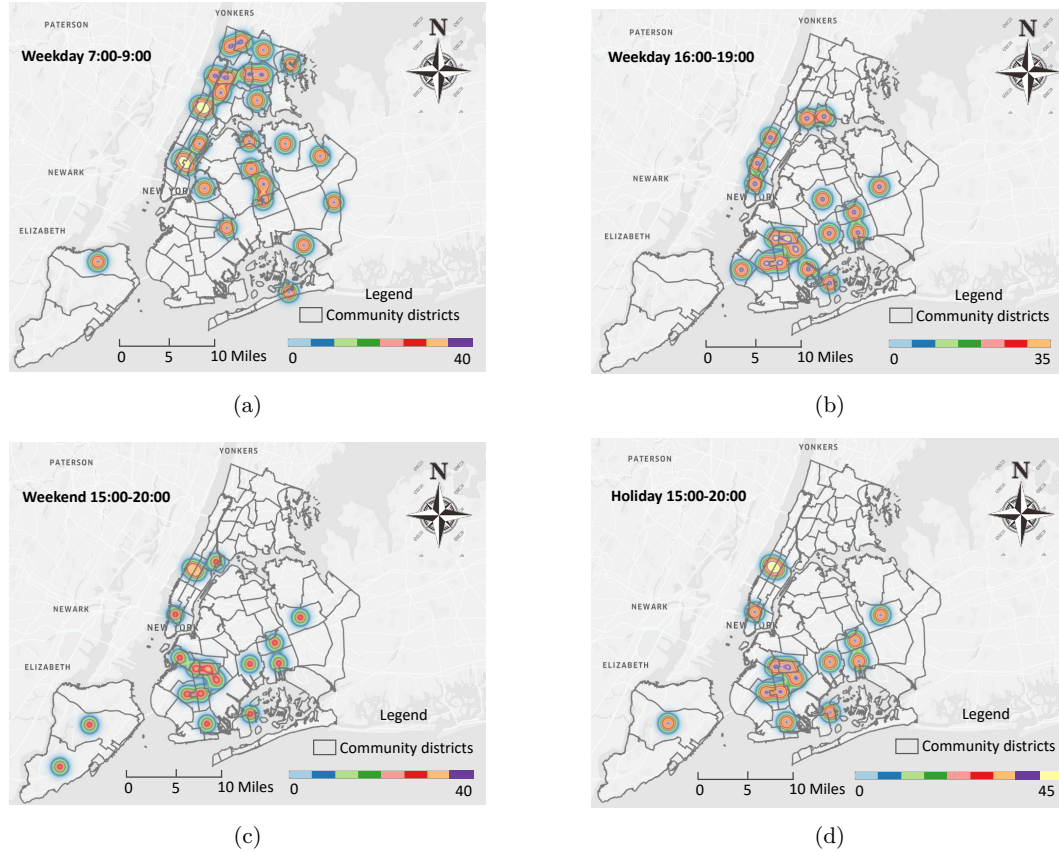


Figure 4 Distribution of congestion communities on different days.

Three significant insights emerge from our findings: First, we observe distinct spatiotemporal asymmetry in congestion patterns. The weekday morning peak (7:00-9:00) and evening peak (16:00-19:00) exhibit contrasting distribution characteristics. Morning congestion demonstrates higher frequency and density concentration in Manhattan, while evening congestion shows a spatial shift towards Brooklyn with greater dispersion. This spatial-temporal variation pattern suggests different commuting behaviors and traffic dynamics between morning and evening rush hours. Second, our analysis reveals temporal stability in congestion center locations, indicating consistent patterns in urban traffic flow distribution across different time periods. This stability has important implications for urban planning and traffic management strategies. Third, the study demonstrates network resilience through the scalability of congestion center identification. The methodology successfully maps road segment-level congestion patterns to community district scales, highlighting its adaptability

across different network resolutions. This multi-scale analytical capability provides valuable insights for urban transportation planning and infrastructure development. These findings contribute to the understanding of urban traffic dynamics by quantifying and visualizing the complex spatiotemporal patterns of congestion centers, offering practical implications for traffic management and urban planning strategies.

5 Conclusion

This study proposes a modularity-driven framework that effectively unravels congestion centers by integrating node degree, spatial proximity, and the fast Fourier Transform of the TSI. Our framework is distinguished by its temporal granularity independence and scalability across multiple spatial scales. The incorporation of FFT-enhanced TSI features significantly improves the model's ability to capture congestion propagation patterns, regardless of the temporal resolution of the input data. This kind of enhanced node feature fusion approach allows to uncover congestion propagation centers regardless of temporal granularity of the datasets. Our approach demonstrated a significant improvement in detecting congestion centers, achieving a modularity score of 0.65 on NYC floating car dataset. The congestion centers identified by our framework offer two advantages over traditional methods: (i) Multi-scale network resilience analysis. Unlike conventional single-scale approaches, our framework leverages road-segment-level data to reveal congestion propagation patterns at broader scales, such as community districts. This multi-scale capability allows for flexible extension to other spatial resolutions, including blocks and boroughs, providing a more comprehensive understanding of urban traffic dynamics. (ii) Dynamic spatiotemporal evolution: Our framework captures the continuous evolution of congestion centers across both time and space, moving beyond static snapshots to provide a more nuanced representation of traffic patterns.

The practical implications of our model and the detected congestion centers are manifold. Firstly, urban planners and authorities can leverage the framework to identify critical congestion hotspots and prioritize infrastructure investments. By understanding the multi-scale resilience of the congestion centers, they can design targeted interventions that reduce bottlenecks. Secondly, The framework's ability to track the dynamic evolution of congestion propagation centers across time and space supports the development of adaptive, real-time traffic management systems. Moreover, the modularity-driven approach of our framework ensures that the detected congestion centers are accurate and interpretable. The interpretability is crucial for stakeholders who need to make informed decisions based on the model's outputs.

6 Declaration of Competing Interest

The author(s) hereby declare that they have no potential conflicts of interest with respect to the research, authorship, or publication of this work. The authors affirm their commitment to maintaining the integrity and objectivity of the research process, ensuring that the work adheres to the highest ethical standards in academic and scientific practice.

7 Data availability

The average hourly speed data in New York City was downloaded from the Uber Movement in March, 2023. But Uber Movement no longer open this data to the public now. The community districts were downloaded from NYC Open data. We are glad to share all above data on request.

References

- 1 Kenneth Li-Minn Ang, Jasmine Kah Phooi Seng, Ericmoore Ngharamike, and Gerald K Ijamaru. Emerging technologies for smart cities' transportation: geo-information, data analytics and machine learning approaches. *ISPRS International Journal of Geo-Information*, 11(2):85, 2022. doi: 10.3390/ijgi11020085. doi:10.3390/IJGI11020085.
- 2 Sanchita Basak, Abhishek Dubey, and Leao Bruno. Analyzing the cascading effect of traffic congestion using lstm networks. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 2144–2153. IEEE, 2019. doi:10.1109/BigData47090.2019.9005995.
- 3 Vincent D Blondel, Jean-Loup Guillaume, Julien M Hendrickx, Cristobald de Kerchove, and Renaud Lambiotte. Local leaders in random networks. *Physical Review E – Statistical, Nonlinear, and Soft Matter Physics*, 77(3):036114, 2008. doi:10.1103/PhysRevE.77.036114.
- 4 Zhenhua Chen, Yongjian Yang, Liping Huang, En Wang, and Dawei Li. Discovering urban traffic congestion propagation patterns with taxi trajectory data. *IEEE Access*, 6:69481–69491, 2018. doi: 10.1109/ACCESS.2018.2881039. doi:10.1109/ACCESS.2018.2881039.
- 5 Xinyue Fan, Jiao Zhang, and Qi Shen. Prediction of road congestion diffusion based on dynamic bayesian networks. In *Journal of Physics: Conference Series*, volume 1176(2), page 022046. IOP Publishing, 2019. doi:10.1088/1742-6596/1176/2/022046.
- 6 Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. doi:10.48550/arXiv.1609.02907.
- 7 Assemgul Kozhabek, Wei Koong Chai, and Ge Zheng. Modeling traffic congestion spreading using a topology-based sir epidemic model. *IEEE Access*, 2024. doi:10.1109/ACCESS.2024.3370474.
- 8 Maohan Liang, Ryan Wen Liu, Yang Zhan, Huanhuan Li, Fenghua Zhu, and Fei-Yue Wang. Fine-grained vessel traffic flow prediction with a spatio-temporal multigraph convolutional network. *IEEE Transactions on Intelligent Transportation Systems*, 23(12):23694–23707, 2022. doi:10.1109/TITS.2022.3199160.
- 9 Sen Luan, Ruimin Ke, Zhou Huang, and Xiaolei Ma. Traffic congestion propagation inference using dynamic bayesian graph convolution network. *Transportation research part C: emerging technologies*, 135:103526, 2022. doi:10.1016/j.trc.2021.103526.
- 10 Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004. doi:10.1103/PhysRevE.69.026113.
- 11 Hoang Nguyen, Wei Liu, and Fang Chen. Discovering congestion propagation patterns in spatio-temporal traffic data. *IEEE Transactions on Big Data*, 3(2):169–180, 2016. doi: 10.1109/TBDATA.2016.2587669. doi:10.1109/TBDATA.2016.2587669.
- 12 Meead Saberi, Homayoun Hamedmoghadam, Mudabber Ashfaq, Seyed Amir Hosseini, Ziyuan Gu, Sajjad Shafiei, Divya J Nair, Vinayak Dixit, Lauren Gardner, S Travis Waller, et al. A simple contagion process describes spreading of traffic jams in urban networks. *Nature communications*, 11(1):1616, 2020.
- 13 Dingyi Shi, Fan Shang, Bingsheng Chen, Paul Expert, Linyuan Lü, H Eugene Stanley, Renaud Lambiotte, Tim S Evans, and Ruiqi Li. Local dominance unveils clusters in networks. *arXiv preprint arXiv:2209.15497*, 2022. arXiv:2209.15497, doi:10.1038/s42005-024-01635-4.
- 14 Shiliang Sun, Changshui Zhang, and Guoqiang Yu. A bayesian network approach to traffic flow forecasting. *IEEE Transactions on intelligent transportation systems*, 7(1):124–132, 2006. doi: 10.1109/TITS.2006.869623. doi:10.1109/TITS.2006.869623.
- 15 Peixiao Wang, Tong Zhang, Yueming Zheng, and Tao Hu. A multi-view bidirectional spatiotemporal graph network for urban traffic flow imputation. *International Journal of Geographical Information Science*, 36(6):1231–1257, 2022. doi:10.1080/13658816.2022.2032081.
- 16 Jianjun Wu, Ziyao Gao, and Huijun Sun. Simulation of traffic congestion with sir model. *Modern Physics Letters B*, 18(30):1537–1542, 2004. doi:10.1142/S0217984904008031.
- 17 Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24, 2020. doi:10.1109/TNNLS.2020.2978386.

- 18 Haoyi Xiong, Xun Zhou, and David A Bennett. Detecting spatiotemporal propagation patterns of traffic congestion from fine-grained vehicle trajectory data. *International Journal of Geographical Information Science*, 37(5):1157–1179, 2023. doi:10.1080/13658816.2023.2178653.
- 19 Kunpeng Zhang, Ning Jia, Liang Zheng, and Zijian Liu. A novel generative adversarial network for estimation of trip travel time distribution with trajectory data. *Transportation Research Part C: Emerging Technologies*, 108:223–244, 2019. doi:10.1016/j.trc.2019.09.019.
- 20 Ling Zhao, Yujiao Song, Chao Zhang, Yu Liu, Pu Wang, Tao Lin, Min Deng, and Haifeng Li. T-gcn: A temporal graph convolutional network for traffic prediction. *IEEE transactions on intelligent transportation systems*, 21(9):3848–3858, 2019. doi:10.1109/TITS.2019.2935152.
- 21 Qi Zheng and Yaying Zhang. Dstagn: Dynamic spatial-temporal adjacent graph convolutional network for traffic forecasting. *IEEE Transactions on Big Data*, 9(1):241–253, 2022. doi:10.1109/TBDATA.2022.3156366.


BERT4Traj: Transformer-Based Trajectory Reconstruction for Sparse Mobility Data

Hao Yang ✉

Department of Geography, University of Georgia, Athens, GA, USA

Angela Yao¹ ✉

Department of Geography, University of Georgia, Athens, GA, USA

Christopher C. Whalen ✉ 

College of Public Health, University of Georgia, Athens, GA, USA

Gengchen Mai ✉

Department of Geography and the Environment, University of Texas at Austin, TX, USA

Abstract

Understanding human mobility is essential for applications in public health, transportation, and urban planning. However, mobility data often suffers from sparsity due to limitations in data collection methods, such as infrequent GPS sampling or call detail record (CDR) data that only capture locations during communication events. To address this challenge, we propose BERT4Traj, a transformer-based model that reconstructs complete mobility trajectories by predicting hidden visits in sparse movement sequences. Inspired by BERT's masked language modeling objective and self-attention mechanisms, BERT4Traj leverages spatial embeddings, temporal embeddings, and contextual background features such as demographics and anchor points. We evaluate BERT4Traj on real-world CDR and GPS datasets collected in Kampala, Uganda, demonstrating that our approach significantly outperforms traditional models such as Markov Chains, KNN, RNNs, and LSTMs. Our results show that BERT4Traj effectively reconstructs detailed and continuous mobility trajectories, enhancing insights into human movement patterns.

2012 ACM Subject Classification Computing methodologies → Machine learning

Keywords and phrases Human Mobility, Trajectory Reconstruction, Deep Learning, CDR, GPS

Digital Object Identifier 10.4230/LIPICs.GIScience.2025.8

1 Introduction

Understanding human mobility is crucial for various applications, including public health, transportation, and urban planning [12, 2, 15]. With the increasing availability of location data from GPS devices, mobile phones, and other portable technologies, human mobility analysis has gained significant attention. However, despite the abundance of location data, data sparsity remains a persistent challenge. For example, Call Detail Records (CDRs) capture locations only when calls or text messages occur, leaving significant gaps in a user's movement trajectory [4]. Similarly, GPS data may be sparse due to battery-saving modes, signal loss, or intermittent sampling. Consequently, there exist places that individuals have visited but are not recorded in the data, which we refer to as “hidden visits” [1]. The presence of hidden visits impedes the ability to reconstruct a complete view of an individual's daily movement, posing substantial challenges to understanding human mobility. Thus, identifying hidden visits to address data sparsity and reconstructing continuous, detailed, and complete mobility trajectories is a necessary and meaningful research problem.

¹ Corresponding Author



© Hao Yang, Angela Yao, Christopher C. Whalen, and Gengchen Mai;
licensed under Creative Commons License CC-BY 4.0

13th International Conference on Geographic Information Science (GIScience 2025).

Editors: Katarzyna Sila-Nowicka, Antoni Moore, David O'Sullivan, Benjamin Adams, and Mark Gahegan;
Article No. 8; pp. 8:1–8:9



Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Early trajectory reconstruction methods primarily relied on Markov Chains and interpolation-based techniques [6, 16, 8, 7]. Markov Chain models, such as the one proposed by [8], incorporate activity changes to enhance mobility prediction but struggle with long-range dependencies and complex movement behaviors due to the adopted Markov assumption. Interpolation-based approaches leverage spatial-temporal correlations to estimate missing points. For instance, [7] use linear and cubic interpolation to reconstruct human mobility from mobile phone data. However, such assumptions fail to capture real-world non-linear travel patterns effectively.

With advancements in machine learning, researchers have increasingly adopted deep learning models for trajectory reconstruction [3, 9, 10, 14]. Backpropagation (BP) neural networks, as proposed by [10], reconstruct mobility trajectories from sparse Call Detail Records (CDR) to estimate hourly population density. However, this method assumes predictable movement patterns, overlooking detours and irregular trajectories.

More recently, Transformer-based approaches have demonstrated superior performance [13, 5]. TrajBERT, introduced by [13], applies BERT-based trajectory recovery with spatial-temporal refinement to address implicit trajectory sparsity. While effective in predicting missing locations, TrajBERT lacks external context modeling, such as user characteristics, dynamic temporal variations, or real-world events, limiting its adaptability for high-accuracy trajectory prediction.

To overcome these challenges, this paper introduces **BERT4Traj**, a novel Transformer-based model for trajectory reconstruction. By leveraging BERT’s bidirectional self-attention mechanism, BERT4Traj effectively captures spatial-temporal dependencies, improving trajectory prediction accuracy. The model is applied to reconstruct complete movement trajectories from both CDR and GPS datasets, demonstrating its robustness in handling data sparsity across different mobility data types. Through context-aware trajectory reconstruction, BERT4Traj enables a more detailed and accurate representation of human mobility patterns, offering valuable insights for applications in public health, urban planning, and transportation analytics.

2 Methodology

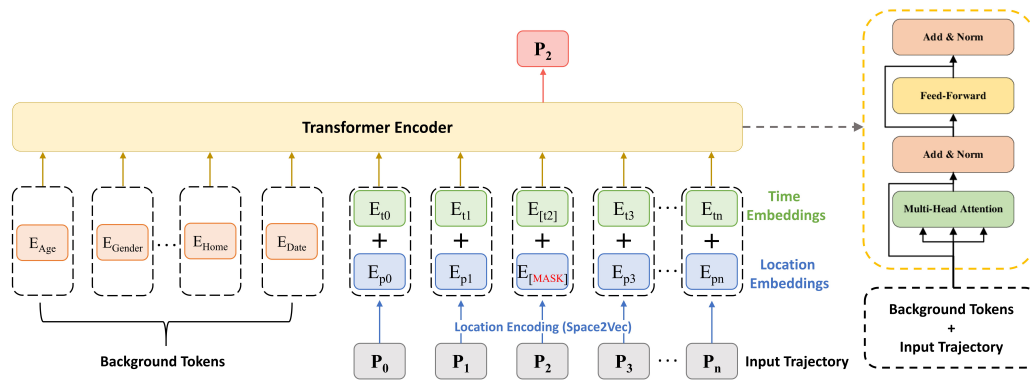
To address the challenge of data sparsity and reconstruct continuous and detailed mobility trajectories, we propose a transformer-based architecture, BERT4Traj, inspired by BERT. This model predicts hidden visits in user mobility trajectories by treating each user’s daily trajectory as a sequence analogous to a sentence in Natural Language Processing (NLP), where locations correspond to ordered words. The objective is to infer missing locations within this sequence using spatial, temporal, and user-specific demographic information.

The core idea of BERT4Traj is inspired by the masked language modeling in BERT, where predictions are made based on contextual information from surrounding tokens. In the context of human mobility, locations visited on the same day provide contextual clues to infer missing visits. In addition to known locations, background information such as demographic characteristics (e.g., age, gender), key life anchors (e.g., home and workplace), and temporal context (e.g., weekday vs. weekend, holidays) further enrich the representation of an individual’s mobility behavior.

As illustrated in Figure 1, BERT4Traj incorporates a BERT-like masking and prediction mechanism. A subset of locations in a trajectory sequence is randomly masked – for example, location P2 in the figure – and the model learns to predict these missing locations using the

context provided by the rest of the sequence. This bidirectional prediction process enables BERT4Traj to develop a deep understanding of how visited locations relate to one another in varying contexts.

The input sequence consists not only of the trajectory data but also of unmasked context tokens that provide additional background information, including temporal attributes, user demographics, and travel characteristics. During training, the model learns intricate relationships between visited locations and contextual features, allowing it to accurately predict missing locations at specific times in a day. Ultimately, this approach reconstructs an individual's movement trajectory with finer temporal granularity, effectively addressing data sparsity issues in mobility datasets.



■ **Figure 1** The overall framework of the BERT4Traj model.

2.1 Data Representation and Embeddings

Each user's daily trajectory consists of a sequence of visited locations with corresponding timestamps. To represent this information in the model, we define location embeddings and time embeddings which are analogy word embeddings and position embeddings in the classic BERT model. Each visited location $\mathbf{x}_i \in \mathbb{R}^2$ is embedded as a vector of dimension d , i.e., $\mathbf{l}_i \in \mathbb{R}^d$. Here, i is the index of the location in the trajectory. The location embeddings encode geographical information (latitude, longitude) and may also include semantic attributes, such as Points of Interest (POI) types or travel modes, if available.

To model temporal dependencies, a time embedding $\mathbf{t}_i \in \mathbb{R}^d$ is generated based on the timestamp capturing the time of the visit. The time embedding functions similarly to positional embeddings in NLP models, providing temporal context to the trajectory sequence.

In addition to trajectory embeddings, we incorporate background tokens representing demographic features, anchor points, and temporal attributes. Specifically, $\mathbf{B} = [\mathbf{w}_{\text{age}}; \mathbf{w}_{\text{gender}}; \dots]$ represents the demographic embeddings, where \mathbf{w}_{age} and $\mathbf{w}_{\text{gender}}$ denote the age and gender embeddings, respectively.

Anchor points, such as home and workplace locations, are encoded as $\mathbf{A} = [\mathbf{w}_{\text{primary}}; \mathbf{w}_{\text{secondary}}; \dots]$, where $\mathbf{w}_{\text{primary}}$ and $\mathbf{w}_{\text{secondary}}$ represent embeddings for primary and secondary anchor points.

Temporal context, including whether the day is a weekday, weekend, or holiday, is represented as $\mathbf{T} = [\mathbf{w}_{\text{weekday}}; \mathbf{w}_{\text{weekend}}; \dots]$, where $\mathbf{w}_{\text{weekday}}$ and $\mathbf{w}_{\text{weekend}}$ denote the corresponding time-related embeddings.

To form the final input to the Transformer encoder, we concatenate these background tokens with the spatiotemporal trajectory tokens. Each trajectory token is constructed by performing element-wise addition of the location embedding \mathbf{l}_i and its corresponding time embedding \mathbf{t}_i , effectively combining spatial and temporal context into a single vector.

The complete input sequence is formulated as:

$$\mathbf{X} = [\mathbf{B}; \mathbf{A}; \mathbf{T}; \mathbf{l}_1 + \mathbf{t}_1; \mathbf{l}_2 + \mathbf{t}_2; \dots; \mathbf{l}_n + \mathbf{t}_n]. \quad (1)$$

2.2 Masking Mechanism

A portion of the location tokens in the trajectory sequence is randomly masked. The objective is to predict these masked locations using unmasked locations and contextual embeddings. Let $\mathbf{M} \in \{0, 1\}^n$ be a binary masking vector, where $\mathbf{M}_i = 1$ if the location \mathbf{l}_i is masked and $\mathbf{M}_i = 0$ otherwise. The masked sequence is represented as:

$$\mathbf{X}_{\text{masked}} = [\mathbf{B}; \mathbf{A}; \mathbf{T}; (\mathbf{M}_1 \cdot \mathbf{l}_1) + \mathbf{t}_1; \dots; (\mathbf{M}_n \cdot \mathbf{l}_n) + \mathbf{t}_n]. \quad (2)$$

This ensures that spatial and temporal relationships are preserved while training the model to infer missing locations.

2.3 Transformer-Based Sequence Encoder

The masked sequence is processed by a Transformer encoder consisting of multiple self-attention layers. The self-attention mechanism computes dependencies between different locations in the trajectory:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (3)$$

where Q, K, V represent the query, key, and value matrices derived from the input sequence, and d_k is the dimensionality of the key vectors.

Multi-head attention further enhances the model's ability to capture diverse mobility patterns:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O. \quad (4)$$

The output of the encoder is a sequence of hidden states $H = [h_1, h_2, \dots, h_n]$, where each h_i encodes contextual information about its corresponding location.

2.4 Masked Location Prediction

In our approach, predicting masked locations is formulated as a classification task rather than a regression problem. This is because the model selects the most likely location from a predefined set of discrete spatial units (e.g., tower IDs for CDR data or grid IDs for GPS data) rather than predicting continuous latitude and longitude values. For each masked location \mathbf{l}_i , the model predicts its most probable value using the output hidden states:

$$\hat{\mathbf{l}}_i = \text{softmax}(Wh_i), \quad (5)$$

where $W \in \mathbb{R}^{|P| \times d}$ is a weight matrix, and $\hat{\mathbf{l}}_i$ represents the predicted probability distribution over the possible locations P . Since each masked location must be assigned one discrete label from a finite set of locations, this naturally aligns with a multi-class classification problem.

The model is optimized using a cross-entropy loss function:

$$\mathcal{L} = - \sum_{i \in \mathcal{M}} \log \hat{\mathbf{l}}_i[\mathbf{l}_i]. \quad (6)$$

Where \mathcal{M} denotes the set of masked locations. Minimizing this loss encourages the model to correctly predict masked locations, improving its ability to reconstruct missing trajectory data.

3 Application of BERT4Traj to CDR and GPS Data

The BERT4Traj framework is highly flexible and can be extended or modified to reconstruct mobility trajectories across different types of mobility data, such as GPS and CDR. Depending on data availability, additional background information can be incorporated to enhance the model's ability to capture mobility behavior and patterns. In this study, to evaluate its effectiveness, we adapted and applied BERT4Traj to two distinct location datasets: Call Detail Records (CDR) collected from 248 cell phone users and GPS data collected from 586 portable watch users in Kampala, Uganda. All participants – both cell phone and watch users – provided self-reported information, including age, gender, anchor points (such as home, workplace, and school locations), education, and income.

3.1 CDR Data

CDR data captures the tower location and timestamp when a communication event occurs, such as a call or text message. Due to its event-driven nature, CDR data is sparse. In our dataset, each individual has records spanning an average of 34 days, but only around five location records per day resulting in significant gaps in their mobility trajectories. To address this, BERT4Traj predicts hidden visits during unrecorded periods, enhancing trajectory completeness.

Each input trajectory is represented as a sequence of tower locations with timestamps. We generate location embeddings using Space2Vec [11], which provides continuous vector representations based on geographical coordinates:

$$e_{loc}(l_i) = \text{Space2Vec}(\mathbf{x}_i), \quad (7)$$

where \mathbf{x}_i denotes the latitude and longitude of the i th tower location.

For time embeddings, we divide the 17-hour time window (from 6:00 AM to 11:00 PM) into 34 half-hour slots, assigning an index from 1 to 34 to each slot. We apply sinusoidal positional encoding to preserve temporal relationships:

$$t_i = \begin{cases} \sin\left(\frac{s_i}{10000^{\frac{2j}{d}}}\right), & \text{if } j \text{ is even} \\ \cos\left(\frac{s_i}{10000^{\frac{2j}{d}}}\right), & \text{if } j \text{ is odd} \end{cases} \quad (8)$$

where s_i is the time slot index (ranging from 1 to 34), j is the embedding dimension index, d is the total embedding dimension. This encoding ensures that nearby time slots have similar embeddings, allowing the model to recognize the temporal structure of the sequence effectively.

Additional context, including age, gender, primary and secondary anchor points, and temporal indicators (e.g., weekday vs. weekend), is incorporated into the model. These background tokens are combined with trajectory tokens and then fed into the Transformer encoder. BERT4Traj predicts tower IDs for the half-hour time slots where no records exist, reconstructing a temporally detailed trajectory.

3.2 GPS Data

GPS data provides higher temporal resolution than CDR but still contains missing records due to device-related issues such as power-saving modes, signal loss, or shutdowns. In our GPS dataset, each individual has an average of 197 days of records, with a mean time interval of 18 minutes between points. To construct continuous trajectories, we apply BERT4Traj to predict the missing locations during these gaps.

The input GPS trajectory consists of exact coordinate points (latitude, longitude) recorded at each timestamp. However, to facilitate prediction and ensure spatial consistency, these exact coordinates are mapped to a $200\text{m} \times 200\text{m}$ grid. Each grid cell has a unique Grid ID, which serves as a spatial unit for the model. To ensure spatial coherence, we derive location embeddings using Space2Vec, following the same approach as with the CDR data.

For time encoding, we use normalized time encoding to preserve the full timestamp (hour, minute) in a continuous and periodic manner. First, we normalize the time to a fraction of the day:

$$t_{\text{norm}} = \frac{\text{hour} \times 60 + \text{minute}}{1440} \quad (9)$$

where 1440 is the total number of minutes in a day.

Then, we apply sinusoidal encoding to capture periodicity:

$$t_i = \begin{cases} \sin(2\pi t_{\text{norm}}) \\ \cos(2\pi t_{\text{norm}}) \end{cases} \quad (10)$$

This encoding ensures that time is represented in a continuous way, maintaining smooth transitions between consecutive timestamps while preserving cyclic properties.

Similarly, we incorporate background tokens representing demographic attributes, anchor points, and temporal indicators. BERT4Traj then reconstructs continuous trajectories by predicting the most likely location (Grid ID) at any given time.

3.3 Model Evaluation

The effectiveness of BERT4Traj was evaluated against several baseline models, including Markov Chain, RNN, LSTM, and KNN, using both CDR and GPS datasets. Performance is assessed using the following metrics:

- **Accuracy:** The proportion of correctly predicted locations.
- **Top-3 Accuracy:** Whether the correct location is among the top three predictions.
- **Top-5 Accuracy:** Whether the correct location is among the top five predictions.

Since the goal of this study is to reconstruct mobility trajectories for known users, we adopt a trajectory-level data split, where each user's daily trajectories are partitioned into 80% for training, 10% for validation, and 10% for testing. For both the CDR and GPS datasets, we apply a random masking strategy during training, where 20% of the location tokens in each trajectory are masked. The model is trained to predict these masked locations using the remaining context and background information. During testing, the same masking

ratio (20%) is applied to the trajectories in the test set. The trained BERT4Traj model is then used to predict the masked locations in these test sequences. Model performance is evaluated by comparing the predicted locations against the true masked labels using metrics such as accuracy, top-3 accuracy, and top-5 accuracy.

Table 1 summarizes the performance comparison. The results clearly demonstrate that BERT4Traj outperforms all baseline models across both datasets. In the CDR dataset, BERT4Traj achieves an accuracy of 87.1%, significantly surpassing LSTM (74.5%) and RNN (70.6%), highlighting its superior ability to handle sparse mobility data compared to recurrent models. Similarly, in the GPS dataset, BERT4Traj attains 71.4% accuracy, outperforming LSTM (62.1%) and RNN (60.3%). The lower accuracy observed in the GPS dataset compared to CDR is due to the fundamental difference in prediction tasks – CDR reconstruction predicts locations within predefined half-hour time slots, whereas GPS trajectory reconstruction requires continuous predictions across time, making the task inherently more challenging. Despite this, BERT4Traj still achieves notable improvements over baseline models, demonstrating its robustness in handling missing data.

Among the baseline models, Markov Chain and KNN exhibit the lowest accuracy, particularly in the GPS dataset, where their accuracy remains below 55%. This indicates that these simpler models struggle to capture sequential dependencies and complex spatial-temporal relationships, reinforcing the advantage of deep learning approaches in trajectory reconstruction.

■ **Table 1** Comparison with baselines in Accuracy, Top-3 Accuracy, and Top-5 Accuracy.

Model	CDR Accuracy (%)	CDR Top-3 (%)	CDR Top-5 (%)	GPS Accuracy (%)	GPS Top-3 (%)	GPS Top-5 (%)
Markov Chain	62.1	65.2	67.8	52.7	53.9	55.3
RNN	70.6	73.4	74.9	60.3	61.2	62.8
LSTM	74.5	77.6	80.1	62.1	64.5	66.4
KNN	67.2	70.4	72.3	54.2	55.4	57.1
BERT4Traj	87.1	89.8	91.2	71.4	73.4	74.8

To examine the contribution of different contextual background features, we conducted an ablation study where demographic information, anchor points, and date information were individually removed from BERT4Traj. The results of this analysis are shown in Table 2 below.

■ **Table 2** Ablation Study: Effect of Removing Contextual Features on Model Accuracy.

Feature Removed	CDR Accuracy (%)	GPS Accuracy (%)
No Demographics	82.7	69.5
No Anchor Points	84.3	71.2
No Date Information	81.5	68.1
Full Model (BERT4Traj)	87.1	71.4

The findings show that removing any contextual feature leads to a decline in model performance. The most significant drop occurs when removing date information, reducing accuracy to 81.5% in the CDR dataset and 68.1% in the GPS dataset. This suggests that temporal context plays a crucial role in predicting missing locations. The removal of demographic data also results in a notable accuracy drop, indicating that user characteristics

contribute valuable information for mobility prediction. Similarly, excluding anchor points reduces accuracy, highlighting their importance in modeling an individual's movement behavior.

Overall, these results demonstrate that incorporating spatial, temporal, and demographic background information enhances the accuracy of BERT4Traj in reconstructing mobility trajectories.

4 Conclusion

This study introduced BERT4Traj, a Transformer-based model for reconstructing complete mobility trajectories from sparse location data. The model effectively captures spatial and temporal dynamic relationships, enabling more accurate trajectory reconstruction. Additionally, BERT4Traj enhances location prediction accuracy by incorporating multi-faceted contextual embeddings, including demographic, anchor point, and temporal features, enriching the representation of human mobility patterns. Moreover, BERT4Traj provides a scalable and adaptable framework for mobility datasets, making it applicable to public health, urban planning, and transportation analytics.

Despite its advantages, BERT4Traj has limitations. Its generalizability across different regions requires further validation, as mobility behaviors vary across geographic and socioeconomic contexts. Privacy concerns also emerge when reconstructing detailed trajectories, necessitating robust safeguards. Future research should explore multi-source mobility data integration, efficiency optimization, and privacy-preserving techniques. Addressing these challenges will enhance BERT4Traj's reliability and applicability in human mobility research and decision-making.

References

- 1 Ian Barnett and Jukka-Pekka Onnela. Inferring mobility measures from gps traces with missing data. *Biostatistics*, 21(2):e98–e112, 2020.
- 2 Vitaly Belik, Theo Geisel, and Dirk Brockmann. Natural human mobility patterns and spatial spread of infectious diseases. *Physical Review X*, 1(1):011001, 2011.
- 3 Cynthia Chen, Ling Bian, and Jingtao Ma. From traces to trajectories: How well can we guess activity locations from mobile phone traces? *Transportation Research Part C: Emerging Technologies*, 46:326–337, 2014.
- 4 Guangshuo Chen, Aline Carneiro Viana, Marco Fiore, and Carlos Sarraute. Complete trajectory reconstruction from sparse mobile phone data. *EPJ Data Science*, 8(1):1–24, 2019. doi:10.1140/EPJDS/S13688-019-0206-8.
- 5 Alessandro Crivellari, Bernd Resch, and Yuhui Shi. Tracebert—a feasibility study on reconstructing spatial-temporal gaps from incomplete motion trajectories via bert training process on discrete location sequences. *Sensors*, 22(4):1682, 2022. doi:10.3390/S22041682.
- 6 Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. Next place prediction using mobility markov chains. In *Proceedings of the first workshop on measurement, privacy, and mobility*, pages 1–6, 2012. doi:10.1145/2181196.2181199.
- 7 Sahar Hoteit, Stefano Secci, Stanislav Sobolevsky, Carlo Ratti, and Guy Pujolle. Estimating human trajectories and hotspots through mobile phone data. *Computer Networks*, 64:296–307, 2014. doi:10.1016/J.COMNET.2014.02.011.
- 8 Wei Huang, Songnian Li, Xintao Liu, and Yifang Ban. Predicting human mobility with activity changes. *International Journal of Geographical Information Science*, 29(9):1569–1587, 2015. doi:10.1080/13658816.2015.1033421.

- 9 Mingxiao Li, Song Gao, Feng Lu, and Hengcai Zhang. Reconstruction of human movement trajectories from large-scale low-frequency mobile phone data. *Computers, Environment and Urban Systems*, 77:101346, 2019. doi:10.1016/J.COMPENVURBSYS.2019.101346.
- 10 Zhang Liu, Ting Ma, Yunyan Du, Tao Pei, Jiawei Yi, and Hui Peng. Mapping hourly dynamics of urban population using trajectories reconstructed from mobile phone records. *Transactions in GIS*, 22(2):494–513, 2018. doi:10.1111/TGIS.12323.
- 11 Gengchen Mai, Krzysztof Janowicz, Bo Yan, Rui Zhu, Ling Cai, and Ni Lao. Multi-scale representation learning for spatial feature distributions using grid cells. In *International Conference on Learning Representations*, 2020.
- 12 Sandro Meloni, Nicola Perra, Alex Arenas, Sergio Gómez, Yamir Moreno, and Alessandro Vespignani. Modeling human mobility responses to the large-scale spreading of infectious diseases. *Scientific reports*, 1(1):62, 2011.
- 13 Junjun Si, Jin Yang, Yang Xiang, Hanqiu Wang, Li Li, Rongqing Zhang, Bo Tu, and Xiangqun Chen. Trajbert: Bert-based trajectory recovery with spatial-temporal refinement for implicit sparse trajectories. *IEEE Transactions on Mobile Computing*, 2023.
- 14 Jingyuan Wang, Ning Wu, Xinxi Lu, Wayne Xin Zhao, and Kai Feng. Deep trajectory recovery with fine-grained calibration using kalman filter. *IEEE Transactions on Knowledge and Data Engineering*, 33(3):921–934, 2019.
- 15 Hao Yang, X Angela Yao, Christopher C Whalen, and Noah Kiwanuka. Exploring human mobility: a time-informed approach to pattern mining and sequence similarity. *International Journal of Geographical Information Science*, pages 1–25, 2024.
- 16 Haofei Yu, Armistead Russell, James Mulholland, and Zhijiong Huang. Using cell phone location to assess misclassification errors in air pollution exposure estimation. *Environmental pollution*, 233:261–266, 2018.

Identifying Resilient Communities in Road Networks: A Path-Based Embedding Approach

Christopher Wagner ✉️🏠^{ID}

Department of Statistics and Applied Probability, University of California, Santa Barbara, CA, USA

Somayeh Dodge ✉️🏠^{ID}

Department of Geography, University of California, Santa Barbara, CA, USA

Danial Alizadeh ✉️🏠^{ID}

Department of Geography, University of California, Santa Barbara, CA, USA

Abstract

Effective resilience analysis of road networks is fundamental to building sustainable and disaster prepared cities. Identifying which road segments share similar vulnerabilities is important for pinpointing high-risk areas within the network and implementing measures to safeguard them against future disruptions. Graph-based community detection can be applied to group together areas of the network sharing similar structural vulnerabilities. However, current graph-based community detection methods either struggle with integrating node features during partitioning or do not account for the path-based dependencies in road networks. This paper introduces the Path-based Community Embedding (PCE) model, an approach that leverages path-based embeddings to overcome these limitations. PCE combines the strengths of graph attention networks and Long Short-Term Memory models (LSTMs) to learn representations that incorporate both local neighborhood information and long-range path dependencies. Our results on the Santa Barbara road network show that PCE improves community detection performance for resilience analysis, thus offering a powerful tool for urban planners and transportation engineers to preemptively identify vulnerabilities in road networks.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases road networks, resilience analysis, machine learning, graph neural networks

Digital Object Identifier 10.4230/LIPIcs.GIScience.2025.9

Funding The authors gratefully acknowledge the financial support from the U.S. National Science Foundation through award #BCS-2043202.

1 Introduction

Road networks serve as the backbone of economic and social development by facilitating the flow of people, goods, and services. With the rapid increase in urbanization, continuous improvement of urban road systems is essential for maintaining the efficiency of transportation infrastructure [11]. A fundamental challenge in transportation engineering is evaluating the resilience of road networks, defined as their inherent capacity to recover performance when faced with disruptive events [16]. Resilience in road networks is crucial for ensuring stable mobility, minimizing economic losses, and enhancing emergency response capabilities in the face of disruptions caused by natural disasters, infrastructure failures, or congestion events [1].

The use of graph theory, where intersections can be represented as nodes and road segments can be represented as edges (or vice versa in a dual graph), has allowed researchers to ask and answer questions revolving around road networks and their resilience [2]. By modeling road networks as graphs, structural properties such as connectivity, centrality [7], and traffic flow can be quantitatively analyzed to assess a network's vulnerability to failures [1]. In particular, community detection in road networks has gained attention as an effective way



© Christopher Wagner, Somayeh Dodge, and Danial Alizadeh;
licensed under Creative Commons License CC-BY 4.0

13th International Conference on Geographic Information Science (GIScience 2025).

Editors: Katarzyna Sila-Nowicka, Antoni Moore, David O'Sullivan, Benjamin Adams, and Mark Gahegan;
Article No. 9; pp. 9:1–9:10



Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

to identify areas that exhibit similar properties [6]. Community detection in graph theory is the process of identifying elements of the network that are closely connected to each other or share similar characteristics. This process reveals the underlying structure of a network by decomposing it into a set of subgroups, called communities or clusters [8]. For resilience analysis, partitioning roads into different communities can provide insights into which areas of the network share structural vulnerabilities or respond similarly to disruptions. During events such as flooding, roads in low-lying areas and near waterways are more susceptible to closures. If these areas can be identified beforehand as part of the same vulnerability community, transportation agencies can implement strategies to allocate resources efficiently to mitigate the effects of a disruption [17].

Traditional community detection methods leverage the structure of the network and partition the nodes into distinct groups by optimizing certain criteria. For instance, modularity maximization methods assign nodes to communities by maximizing the density of connections within the group compared to a random baseline [18]. Spectral clustering uses the eigenvalues of the graph's Laplacian to identify communities by minimizing the number of cuts between groups [20]. Hierarchical clustering recursively divides nodes according to their connectivity patterns to form a tree structure that reveals different levels of the community [6]. Although effective, these approaches often struggle to incorporate information from nodes such as traffic flow dynamics, road capacity, historical data of disturbance and spatial dependencies [1].

With the rise of big data, modern approaches have leveraged the power of machine learning models on graphs to overcome these limitations by learning embeddings that encode both structural information and node (or edge) features. Models such as graph convolutional networks (GCNs) [13], graph attention networks (GATs) [21], and graph autoencoders (GAEs) [12] have enabled more adaptive community detection by integrating node and edge attributes along with temporal patterns. Additionally, machine learning models tend to be more flexible than traditional models, which rely on fixed assumptions about network topology. For example, spectral clustering implicitly assumes that the cluster structure is encoded in the leading eigenvectors of the graph's Laplacian, which holds when the Laplacian has a few small eigenvalues corresponding to well-separated communities [20]. Unlike traditional approaches, machine learning models can identify patterns in how disruptions affect different segments of a road network without many assumptions. Leveraging these methods can help identify communities that not only reflect connectivity patterns, but also resilience related properties (e.g. vulnerability, centrality, proximity to fire, distance to flood risk areas, etc.) in a data-adaptive manner. While these advancements have improved adaptive community detection, they mainly focus on local connectivity patterns and typically overlook the broader structural dependencies that influence network behavior during disruptions. Incorporating path-based embeddings addresses this limitation by capturing sequences of interconnected nodes rather than neighborhoods to create a more comprehensive representation of road network topology.

In this paper, we highlight the role of machine learning, particularly graph-based neural architectures, in road network community detection to quantify disruption resilience. Additionally, we introduce a model that accounts for path-based dependencies in order to reveal structural patterns linking roads with similar resilience characteristics. We evaluate our approach on a real world road network in Santa Barbara, California to assess its effectiveness in community detection compared to the baseline methods.

The rest of this paper is structured as follows: section 2 describes our methodology, including data preprocessing, the embedding generation, and the proposed clustering approach; section 4 presents experimental results, the analysis, and the discussion; and section 5 discusses conclusions and future directions.

2 Methods

This section introduces our proposed Path-based Community Embedding (PCE) model and formalizes the problem of generating embeddings for community detection.

Given a road network represented as a graph $G = (V, E)$, where $V = \{v_1, v_2, \dots, v_n\}$ is the set of nodes (i.e., intersections) and $E \subseteq V \times V$ is the set of edges (i.e., road segments), the goal is to learn node embeddings that facilitate community detection. A community is defined as a set of nodes that exhibit strong structural and functional similarity, which we infer from the graph-based embeddings and agglomerative clustering. Each node v_i has a feature vector $\mathbf{x}_i \in \mathbb{R}^d$, and the network structure is captured by an adjacency matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, where $A_{ij} = 1$, if $(v_i, v_j) \in E$, otherwise $A_{ij} = 0$.

The proposed model consists of three main components: a Graph Attention Network (GAT) encoder [21], a path-based LSTM embedding module [10], and a reconstruction-based decoder. The GAT encoder generates node embeddings by aggregating features from neighbors with learned attention weights:

$$\mathbf{z}_i = \sigma \left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij} \mathbf{W} \mathbf{x}_j \right), \quad i \in V \quad (1)$$

where \mathbf{z}_i represents the learned embedding for node i , which is computed by aggregating information from its neighbors in $\mathcal{N}(i)$, the term \mathbf{x}_j is the input feature vector of node j , and \mathbf{W} is a trainable weight matrix that transforms the input features into a new feature space. The function $\sigma(\cdot)$, the ReLU activation function, is applied element-wise to introduce non-linearity in the learned embeddings. The attention coefficient α_{ij} represents the importance weight assigned to the feature vector of the neighbor node j when aggregating information for the node i , which is defined as [21]:

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(\mathbf{a}^T [\mathbf{W} \mathbf{x}_i \parallel \mathbf{W} \mathbf{x}_j]))}{\sum_{k \in \mathcal{N}(i)} \exp(\text{LeakyReLU}(\mathbf{a}^T [\mathbf{W} \mathbf{x}_i \parallel \mathbf{W} \mathbf{x}_k]))} \quad (2)$$

To capture long-range dependencies, paths $P = \{p_1, p_2, \dots, p_m\}$ are sampled from the network, where each p_l is a sequence of nodes. The embeddings along a path are aggregated as:

$$\mathbf{z}_{p_j} = \frac{1}{|p_j|} \sum_{v \in p_j} \mathbf{z}_v, \quad j = 1, \dots, m \quad (3)$$

These path embeddings are processed using a bidirectional LSTM to extract sequential dependencies, producing refined path embeddings \mathbf{h}_{p_l} . The final node embeddings are then obtained by aggregating the node embeddings and the path embeddings that the node participates in:

$$\mathbf{z}_i^{\text{final}} = \mathbf{z}_i + \frac{1}{|\mathcal{P}(i)|} \sum_{p_l \in \mathcal{P}(i)} \mathbf{h}_{p_l} \mathbb{I}(i \in p_l) \quad (4)$$

where $\mathcal{P}(i)$ is the set of paths that node i belongs to and $\mathbb{I}(i \in p_l)$ is the indicator function that equals 1 if node i is in path p_l and 0 otherwise.

To ensure the embeddings capture meaningful structure, a decoder reconstructs the original node features using a linear transformation followed by a non-linear activation function. This maps the learned embeddings back to the input feature space. To ensure the community structure is learned, the model is trained in an unsupervised manner using a

joint loss function that combines reconstruction and disruption risk contrastive components. Let \mathbf{x}_i denote the input features for node i and $\hat{\mathbf{x}}_i$ the reconstructed features. Next, we compute a composite risk score r_i for each node using normalized signals: betweenness centrality \tilde{b}_i , distance to fire boundaries $\tilde{d}_{\text{fire},i}$, and distance to flood zones $\tilde{d}_{\text{flood},i}$. Specifically, $r_i = w_1 \cdot \tilde{b}_i + w_2 \cdot (1 - \tilde{d}_{\text{fire},i}) + w_3 \cdot (1 - \tilde{d}_{\text{flood},i})$, where w_1 , w_2 , and w_3 are scalar weights. We define a contrastive loss over pairs of nodes (i, j) where $|r_i - r_j| > \delta$ which is given by:

$$\mathcal{L}_{\text{total}} = \alpha \cdot \frac{1}{N} \sum_{i=1}^N \|\hat{\mathbf{x}}_i - \mathbf{x}_i\|_2^2 + \beta \cdot \frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} \max(0, \cos(\mathbf{z}_i, \mathbf{z}_j) - \tau) \quad (5)$$

where \mathcal{P} is the set of high risk difference node pairs and α , β , and τ are hyperparameters controlling the tradeoff between reconstruction and risk aware separation.

After training, agglomerative clustering is applied to $\mathbf{z}^{i^{\text{final}}}_i \in V$ to detect communities. This is a hierarchical clustering method which iteratively merges the most similar node embeddings based on the Euclidean distance and Ward’s linkage criterion [22], which minimizes the variance within each cluster at every merging step. This clustering method helps preserve spatial continuity while allowing for flexible determination of the number of communities based on the road network topology.

3 Experiments

To evaluate the proposed model, we use a road network representing the downtown Santa Barbara area in California. The city of Santa Barbara is a south-facing coastal town, situated between the Santa Ynez Mountains and the Pacific Ocean. The downtown area, depicted in Figure 1 is located in the lower coastal plain where there is a historic flood risk (shown in blue), while the higher grounds in the foothills are closer to chaparral and forest areas with heightened wildfire risk based on historical fires (shown in orange). The road network is typical of a coastal city in California with restricted geography, featuring Interstate Highway 101 running through the city and serving as a major transportation corridor. The highway closely parallels the coastline, connecting the downtown area to nearby regions while navigating the narrow space between the mountains and the ocean. The road network data, derived from OpenStreetMap (OSM) using OSMnx package [4], consists of 2105 nodes (intersections) and 4234 edges (roads), includes all road types, from major highways to residential streets. For each road segment (edge), we incorporated several features relevant to community resilience and vulnerability, including: proximity to previous fire perimeter boundaries, proximity to historical flood risk zones, betweenness centrality (measuring the importance of a road segment for network flow), closeness centrality (measuring the accessibility of a road segment to all other segments), degree centrality (the number of connections a road segment has). These features were chosen to capture both the topological characteristics of the road network and its exposure to various hazards. The betweenness, closeness, and degree centrality measure were each calculated using the formulas as specified in [19] and are implemented using the NetworkX Python package [9].

The proximity values are obtained by measuring the shortest Euclidean distance from each intersection (node) in the road network to the nearest boundary of the fire perimeter or flood risk zone. Figure 1 presents a map of these boundaries in the city of Santa Barbara. Each intersection is treated as a single point and the distance was computed to the closest edge of the respective hazard polygon. The fire and flood data were obtained from the Santa Barbara County historical fire database and the Federal Emergency Management Agency (FEMA) 100 and 500 year flood risk zones for Santa Barbara, which were both accessed from DataBasin [5].

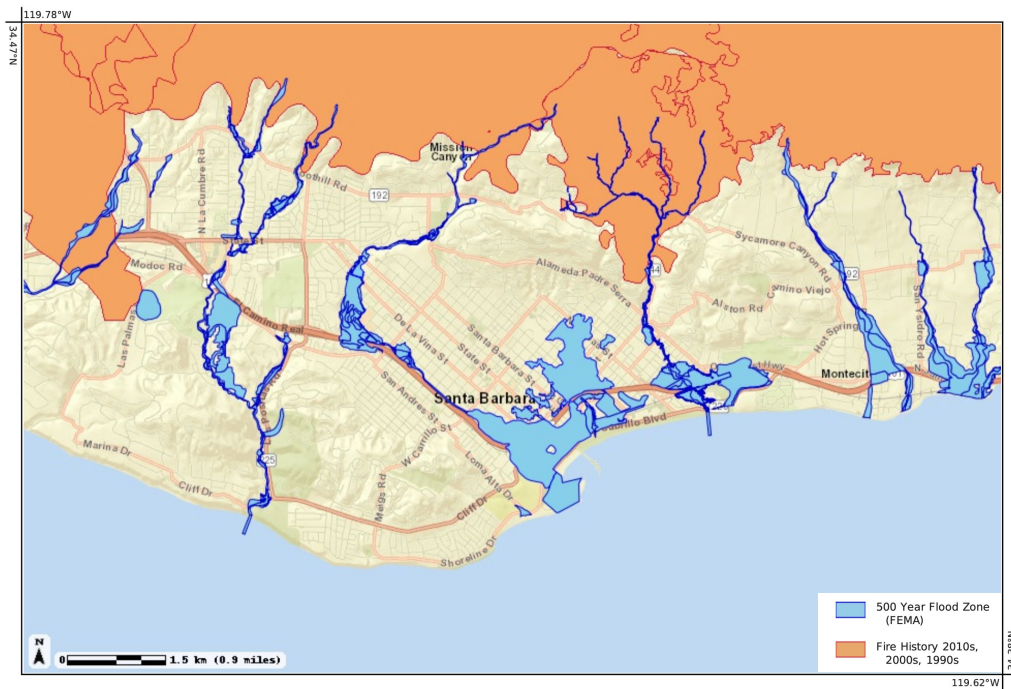


Figure 1 Downtown Santa Barbara with historic fire and flood risk zones, generated from DataBasin.

To evaluate the performance of our model, we compare it against several community detection models. First, the K-means clustering, which partitions nodes into different clusters by minimizing intra-cluster variance in the feature space [15]. It is noted that the K-means clustering method does not consider the graph structure nor does it produce embeddings. The Louvain algorithm is a hierarchical method that iteratively merges nodes into communities to maximize modularity [3]. This method does not require specifying the number of clusters in advance, thus we tune the resolution parameter which controls the granularity of the detected communities. Region2Vec is a spatially-aware graph embedding method that incorporates both node attributes and spatial interactions to detect communities in spatial networks [14]. It uses Graph Convolutional Networks (GCNs) to generate embeddings that balance structural connectivity and spatial proximity before applying clustering to detect regions with similar properties. Graph Autoencoder (GAE) is an unsupervised model that learns node embeddings by reconstructing the graph's adjacency matrix using a GCN [12]. The Spatial Graph Autoencoder (SGAE) extends GAE by simply weighting neighboring nodes based on their spatial proximity, which helps enforce spatial contiguity in the learned embeddings. For each model, we chose the number of clusters to be four for all models except Louvain, since pre-defining this hyperparameter is not supported. This number of clusters provides a good balance between interpretability and meaningful distinctions in the road network's structure.

We compare the models using three key metrics following the convention used in [14]. The first is cosine similarity, which measures the angular similarity between two node embeddings. The resulting value ranges from -1 to 1 , where 1 indicates identical vectors (perfect similarity), 0 indicates orthogonality (no similarity), and -1 indicates completely opposite vectors. Second, Join Count Ratio (JCR) [14] measures the proportion of edges within the road network that connect nodes belonging to the same community. Given a graph $G = (V, E)$ and a community assignment function $C : V \rightarrow \{1, 2, \dots, K\}$, we define:

$$\text{JCR} = \frac{J_{\text{same}}}{J_{\text{same}} + J_{\text{diff}}}, \quad (6)$$

where J_{same} is the number of edges $(u, v) \in E$ where $C(u) = C(v)$, and J_{diff} is the number of edges where $C(u) \neq C(v)$. A higher JCR indicates that the detected communities are more spatially contiguous. Third, Modularity (Q) [18] evaluates the strength of the community structure by comparing the fraction of edges within detected communities to the expected fraction in a random graph with the same degree distribution. It is defined as:

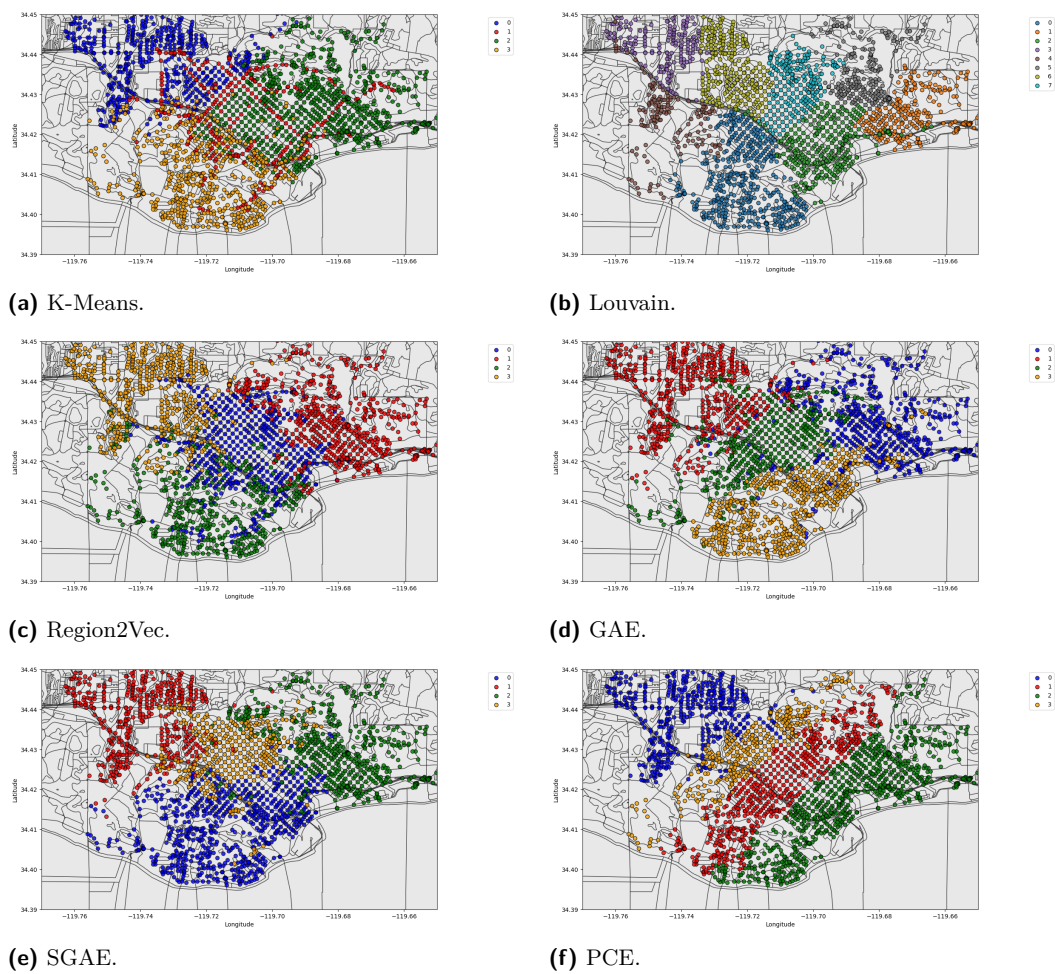
$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(C_i, C_j), \quad (7)$$

where A_{ij} is the adjacency matrix of the graph, k_i and k_j are the degrees of nodes i and j , m is the total number of edges in the graph, and $\delta(C_i, C_j)$ is an indicator function that equals 1 if nodes i and j belong to the same community and 0 otherwise. Higher modularity values correspond to structurally cohesive communities, since they show a greater presence of intra-community edges compared to inter-community edges. These metrics allow us to assess the quality of the embeddings along with the spatial contiguity and structural strength of the predicted communities.

4 Results and Discussion

Figure 2 presents the detected communities using our proposed Path-based (PCE) Community Embedding model, compared to other baseline community detection models: K-Means, Louvain, Region2Vec, Graph Autoencoder (GAE). Table 1 presents a quantitative comparison across the three metrics described in Section 3. The results indicate that the PCE model outperforms most baselines across the three evaluation metrics which can highlight its usefulness in capturing community structures in the road network. It exhibits the highest cosine similarity, which means the embeddings effectively capture the similarity structure of nodes. Additionally, it achieves the highest modularity score and the second highest Join Count Ratio, hence the detected communities form strongly connected groups while maintaining spatial contiguity. The overall results from the community detection across different methods is shown in Figure 2.

The Louvain method, a modularity-based community detection algorithm, achieves the highest JCR. Unlike other models, it relies solely on network topology and excels at preserving connectivity, which may cause it to miss higher-level feature similarities. Thus, it is highly effective at grouping nodes into spatially contiguous regions. Interestingly, although the Louvain method is designed to maximize modularity, the GAE, SGAE, and PCE models achieve higher modularity. Region2Vec and the Graph Autoencoder (GAE) have strong performance across all metrics. Region2Vec, trained to capture spatial proximity and network structure, performs particularly well in cosine similarity and JCR. The K-Means algorithm has the lowest performance across all metrics, particularly modularity and join count ratio.



■ **Figure 2** Results from all community detection methods for the Santa Barbara road network.

■ **Table 1** Comparison of community detection methods. Highest values are bolded.

Method	Cosine Similarity	Join Count Ratio	Modularity
K-Means	—	0.7614	0.3825
Louvain	—	0.9468	0.6060
Region2Vec	0.9314	0.8559	0.5972
GAE	0.8908	0.8669	0.6152
SGAE	0.9820	0.9374	0.6478
PCE	0.9883	0.9432	0.6641

This highlights an expected limitation of feature-space clustering methods applied to road networks: they often disregard the underlying graph structure. Because K-Means ignores connectivity constraints, it may assign distant nodes to the same cluster based on feature similarity, leading to fragmented and spatially disjoint communities. This is reflected in its results, which highlight that clustering methods that overlook graph topology are not well-suited for road network community detection.

PCE’s strong performance can be attributed to its hybrid approach. Its GAT encoder allows nodes to selectively aggregate information from relevant neighbors, while the LSTM-based path encoder captures long-range dependencies in order to ensure that detected communities align with observed connectivity patterns. The integration of neighborhood-based and path-based embeddings ensures that communities are both spatially contiguous and structurally meaningful (respecting graph topology). PCE’s high modularity suggest that it detects highly interconnected subgraphs, which are less vulnerable to isolated disruptions. This is ideal in the context of disruption resilience where communities should represent network regions that can maintain connectivity and functionality during disruptions. In applications like traffic flow management where it’s important to ensure that disruptions in one region don’t severely impact connected areas, this model and other machine learning approaches can potentially help identify regions needing reinforcement to maintain network stability.

Furthermore, the physical and spatial characteristics of the detected communities (Figure 2) offer valuable insights into their resilience profiles. In the visualization of the clusters obtained from PCE in Figure 2f, the densely connected Cluster 3 (yellow) is located in a suburban area with inherent capacity to absorb localized disruptions due to its strong internal connectivity. Conversely, the more dispersed Cluster 0 (blue) shows a community potentially more vulnerable to fire disruptions along critical connecting routes, despite its geographical spread. The core of the network represented by Cluster 1 (red) has high internal connectivity since it contains the roads with the highest centrality, which can provide alternative routing strategies to prevent cascading failures in the event of major disruptions. Finally, Cluster 2 (green) is the one situated closest to the water and the flood risk zones which may mean increased flood risk vulnerability. These clusters demonstrate PCE’s ability to not only identify community structures but also reveal information about their strengths and weaknesses under disruptive events. Our findings suggest that our model can effectively capture the underlying structural characteristics that contribute to resilience in road networks which is valuable for targeted interventions and resilience planning.

5 Conclusion

In this paper, we tackled the challenging problem of community detection in road networks, focusing specifically on identifying communities with shared resilience characteristics. We introduced a novel graph-based embedding model that effectively captures both local and global structural information within the network. Our model PCE leverages Graph Attention Networks to learn local patterns and path-based LSTM to learn long-range global dependencies. This allows the model to understand how nodes are connected across the network, which captures broader structural relationships that are crucial for identifying resilient communities. The combined embeddings that incorporate both local and global perspectives are then used to reconstruct the original node features in a self-supervised manner. Finally, we employ agglomerative clustering on these learned embeddings to reveal the community structure.

Our key contribution lies in the unique combination of local and global graph information, enabling the identification of communities that are not only spatially contiguous but also share resilience properties due to their structural organization. We demonstrated the effectiveness of our approach on the Santa Barbara road network, where we were able to identify distinct communities. These findings suggest that our model can effectively capture the underlying structural characteristics that contribute to resilience in road networks. Furthermore, the unsupervised nature of our approach makes it applicable to a wide range of road network analysis tasks, even when labeled data is scarce or unavailable.

There are several limitations to this study, the main one being the computational complexity of the model and the calculated features, namely betweenness and closeness centrality. As the size of the network increases (e.g. over 100,000 nodes), the computation time will skyrocket and thus may be infeasible. Future work could explore ways to decrease the time complexity of computing these metrics and the model itself to enhance the scalability and assess the generalizability of the model. The model can also be further extended to incorporate dynamic network information (e.g, real-time movement flows) to further enhance its ability to identify resilient communities.

References

- 1 Danial Alizadeh and Somayeh Dodge. Disaster vulnerability in road networks: a data-driven approach through analyzing network topology and movement activity. *International Journal of Geographical Information Science*, pages 1–22, 2024. doi:10.1080/13658816.2024.2411001.
- 2 Chandra Balijepalli and Raphael Oppong. Measuring vulnerability of road network considering the extent of serviceability of critical road links in urban areas. *Journal of Transport Geography*, 46:65–75, 2015.
- 3 Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008. doi:10.1088/1742-5468/2008/10/P10008.
- 4 Geoff Boeing. Modeling and analyzing urban networks and amenities with osmnx. *Geographical Analysis*, 2025.
- 5 Santa Barbara County Fire Department and CAL FIRE. Fire history, santa barbara county, 1990-2020, 2021. Accessed: 2025-02-19. URL: <https://databasin.org/datasets/322c741a26d24400a1ab948d40887fad/>.
- 6 Santo Fortunato and Darko Hric. Community detection in networks: A user guide. *Physics Reports*, 659:1–44, 2016.
- 7 Linton C Freeman et al. Centrality in social networks: Conceptual clarification. *Social network: critical concepts in sociology*. Londres: Routledge, 1(3):238–263, 2002.
- 8 Michelle Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002. doi:10.1073/pnas.122653799.
- 9 Aric Hagberg, Daniel Schult, and Pieter Swart. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conference (SciPy)*, pages 11–15, 2008.
- 10 Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. doi:10.1162/neco.1997.9.8.1735.
- 11 Yang Hong and Yao Yao. Hierarchical community detection and functional area identification with osm roads and complex graph theory. *International Journal of Geographical Information Science*, 33(8):1569–1587, 2019. doi:10.1080/13658816.2019.1584806.
- 12 Thomas N Kipf and Max Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016. arXiv:1611.07308.
- 13 Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations (ICLR)*, 2017. arXiv:1609.02907.
- 14 Yunlei Liang, Jiawei Zhu, Wen Ye, and Song Gao. Geoai-enhanced community detection on spatial networks with graph deep learning. *Computers, Environment and Urban Systems*, 117:102228, 2025. doi:10.1016/j.compenvurbsys.2024.102228.
- 15 J. MacQueen. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

- 16 B. Martín, E. Ortega, R. Cuevas-Wizner, A. Ledda, and A. De Montis. Assessing road network resilience: an accessibility comparative analysis. *Transportation Research Part D: Transport and Environment*, 95:102851, 2021. doi:10.1016/j.trd.2021.102851.
- 17 Lars-Göran Mattsson and Erik Jenelius. Vulnerability and resilience of transport systems—a discussion of recent research. *Transportation Research Part A: Policy and Practice*, 81:16–34, 2015.
- 18 M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.
- 19 M. E. J. Newman. *Networks: An Introduction*. Oxford University Press, 2010.
- 20 Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 849–856, 2001. URL: <https://proceedings.neurips.cc/paper/2001/hash/801272ee79cfde7fa5960571fee36b9b-Abstract.html>.
- 21 Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *International Conference on Learning Representations (ICLR)*, 2018.
- 22 Joe H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963. doi:10.1080/01621459.1963.10500845.

Geovicla: Automated Classification of Interactive Web-Based Geovisualizations

Phil Hüffer ✉ 

Institute for Geoinformatics, University of Münster, Germany

Auriol Degbelo¹ ✉ 

Chair of Geoinformatics, TU Dresden, Germany

Benjamin Risse ✉ 

Institute for Geoinformatics, University of Münster, Germany

Abstract

The exponential growth of interactive geovisualizations on the Web has underscored the need for automated techniques to enhance their findability. In this paper, we present the *Geovicla* dataset (2.5K instances), constructed through the harvesting and manual labelling of webpages from a broad range of domains. The webpages are categorized into three groups: “interactive visualisation”, “interactive geovisualisation” and “no interactive visualisation”. Using this dataset, we compared three approaches for interactive (geo)visualization classification: (i) a heuristic-based approach (i.e. using manually derived rules), (ii) a feature-engineering approach (i.e. hand-crafted feature vectors combined with machine learning classifiers) and (iii) an embedding-based approach (i.e. automatically generated large language model (LLM) embeddings with machine learning classifiers). The results indicate that LLM embeddings, when used in conjunction with a multilayer perceptron, form a promising combination, achieving up to 74% accuracy for multiclass classification and 75% for binary classification. The dataset and the insights gained from our empirical comparison offer valuable resources for GIScience researchers aiming to enhance the discoverability of interactive geovisualizations.

2012 ACM Subject Classification Human-centered computing → Geographic visualization; Information systems → Web searching and information discovery; Information systems → Specialized information retrieval

Keywords and phrases spatial information search, geovisualization search, findable interactive geovisualization, webpage classification

Digital Object Identifier 10.4230/LIPICs.GIScience.2025.10

Supplementary Material *Software*: <https://github.com/phuef/ma/> [20]

Dataset: <https://doi.org/10.34740/kaggle/dsv/10703824>

Funding *Auriol Degbelo*: Auriol Degbelo is funded by the German Research Foundation through the project NFDI4Earth (DFG project no. 460036893, <https://www.nfdi4earth.de/>) within the German National Research Data Infrastructure (NFDI, <https://www.nfdi.de/>).

1 Introduction

Interactive visualisations are becoming increasingly available on the Web and techniques are needed to facilitate their findability [11]. Since maps are “one of the most valuable document for gathering geospatial information about a region” [17], finding and accessing this type of data is relevant for tasks such as information synthesis and hypothesis generation about places during the early phases of the research data lifecycle. Currently, finding interactive

¹ Corresponding author



© Phil Hüffer, Auriol Degbelo, and Benjamin Risse;
licensed under Creative Commons License CC-BY 4.0

13th International Conference on Geographic Information Science (GIScience 2025).

Editors: Katarzyna Sila-Nowicka, Antoni Moore, David O’Sullivan, Benjamin Adams, and Mark Gahegan;
Article No. 10; pp. 10:1–10:12



Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

maps for specific tasks remains challenging, though there are some solutions – in the form of online platforms – that offer limited cataloging functionalities (e.g. Observable [30] and ArcGIS Online Gallery [15]).

The focus of this work is on the automated classification of interactive geovisualizations of the Web. While different approaches to classifying webpages have been proposed in the literature (see [7, 19, 31, 41] for examples and [9, 34] for reviews), the categorization of interactive visualization and interactive geovisualization has, so far, received less attention. Interactive (geo)visualization classification can be seen as an instance of *genre* classification, which, as discussed in [9], is about categorizing webpages based on functional factors, unlike subject-based classification that focuses on their topic. In general, the practical relevance of automated classification of resources in the context of spatial information search is at least twofold: resource selection [8, 10] and results presentation [28] (e.g. in the form of structured and actionable results).

“Resource selection” is a task in distributed information retrieval (a.k.a federated search), which consists in finding the most relevant data sources for a user’s query in a heterogeneous collection. Resource selection, in this context, has the potential to improve users’ satisfaction during interactive (geo)visualization search through the identification of the most related types of target entities to their search intent. This is particularly relevant in the context of scientific [4] and spatial data infrastructures [13], which feature heterogeneous collections of (geoinformation) resources. Besides, the identification of the type of search targets is key to structured result presentation and actionable results presentation.

“Structured results presentation” and “actionable results presentation” are two patterns for the design of search user interfaces, as discussed in [28]. Both approaches enable users to access the information they need without having to open complete result pages. “Structured results presentation” is concerned with using rich snippets (e.g. maps, timelines) to *communicate the structure* of search results (e.g. spatial structure, temporal structure) in addition to simple text snippets (e.g. title, description). “Actionable results presentation” involves providing the means to *perform tasks* as an integral part of the result presentation process (e.g. zooming/panning an interactive map, playing/stopping an animated geovisualization).

This article presents an exploratory study that addresses the research question: Which classification methods are best suited for identifying webpages containing interactive geovisualisations? In line with Koehler [24], webpages are defined throughout this article as collections of Internet objects navigable without hypertext links; they are web documents that can be scrolled through. Websites consist of one or more webpages unified by a common theme or organizing principle. The contributions of the work are twofold: First, we present the *Geovicla* dataset, which was constructed through the harvesting and manual labelling of webpages from a broad range of domains (e.g. sustainability, health, technology, human rights and politics). The dataset includes 2.5K annotated webpage instances from diverse domains and provides labels for three categories, namely “interactive visualisation” (IV), “interactive geovisualisation” (IGV) and “no interactive visualisation” (noIV).

Second, we compared three approaches for the automated interactive (geo)visualizations classification: (i) a heuristic-based approach, (ii) a feature-engineering approach and (iii) an embedding-based approach. Approach (i) uses manually derived rules and heuristics to identify IV and IGV based on the webpages’ code; approach (ii) utilises hand-crafted feature vectors in combination with a machine learning classifier and approach (iii) automatically extracts embeddings using a large language model (LLM), which are subsequently used to classify the web content.

2 Background

The focus of the article is on web-based interactive (geo)visualizations, which at their core, are web documents as discussed in [11]. Here, we briefly touch upon previous work on static map search and classification, as well as interactive map search and classification.

Regarding *Static Map Search and Classification*, existing approaches have tackled the issue from different perspectives, often with very different goals. For example, Goel et al. [17] used a Content-Based Image Retrieval (CBIR) approach to classify static images extracted from PDF files and the Web as maps or nonmaps, achieving an F1 score of 74%. Tan et al. [36] investigated the classification of figures in digital documents as maps or nonmaps and used several variants of support vector machines (SVM) for the classification task. They reported F1 measures of up to 90%. While the two articles mentioned above have a stronger focus on image classification in digital documents, others emphasize Web image harvesting and classification. For instance, Beagle [3] mines the Web for SVG-based (Scalable Vector Graphics) visualizations and automatically classifies them by type (e.g. bar charts, line charts, maps, ...). The authors reported an accuracy of 85% across 24 visualization types. Bone et al. [5] proposed a Geospatial Search Engine that harvests Web Map Services and ArcGIS services (among others) to provide enhanced searchability. Finally, Walter et al. [38] tested several approaches to automate the harvesting of maps in the shapefile format on the Web. They found that the combination of a crawler and a search engine is more efficient than the use of a crawler alone and reported a hit rate during search between 0.18% and 1.5%. We use a search engine during our harvesting workflow in line with this finding (Section 3).

Concerning *Interactive Map Search and Classification*, a research agenda for findable online geovisualization was proposed in [11], highlighting three aspects: knowledge representation aspects, user interface design issues, and technical considerations during the publishing of online geovisualizations. Previous work has focused on user interaction aspects and publishing aspects mostly. For example, Degbelo et al. [12] examined design elements for the search of map layers in map-based applications, while Hüffer et al. [21] compiled users' wishes regarding search tools for interactive (geo) visualizations through participant interviews. Regarding the publishing of online geovisualizations, Lai and Degbelo [26] compared the impact of speech-based and typing modalities for the creation of metadata for web maps and provided empirical evidence about their complementarity for effective geovisualization annotation. Thompson et al. [37] proposed the MIAGIS standard to facilitate the publication of maps according to the FAIR principles and illustrated how the standard can be used to publish maps generated within ArcGIS Online. We argue here that while these works are valuable, progress regarding knowledge representation is equally important to advance current research on findable online geovisualizations. Classification, i.e. finding the semantic type (a.k.a. category) of web documents, is a key aspect of knowledge representation and is the subject of this article.

3 The Geovicla Dataset

Open datasets about interactive (geo)visualizations are desirable to advance research on interactive (geo)visualization search but are still lacking. The generation of the *Geovicla* dataset to address this gap considered the following three categories of web documents.

Interactive visualisation (IV): An interactive visualisation is a webpage, which displays at least one visualisation that affords computer-mediated interaction. Interaction in this context is defined in line with [14, 35] as the dialogue, involving a data-related intent, between a human and a data interface.

Interactive geovisualisation (IGV): An interactive geovisualisation is a webpage, which shows at least one geovisualization that affords computer-mediated interaction. Interaction is defined as stated above; a geovisualization is a digital artefact whose visual properties encode geographic data [11].

No interactive visualisation (noIV): This category is used to refer to webpages that do not contain an IV or IGV, as defined above.

The generation of the dataset involved three tasks, namely search term generation, web document search and web document labelling.

Search term generation: To generate search queries with a high possibility of finding interactive visualisations and interactive geovisualisations, we employed the commonly available ChatGPT model [32], with GPT version 3.5. In particular, this model was used to generate synonyms for the phrases “interactive visualisation” and “interactive geovisualisation”, as well as a set of random topics to query for webpages. Examples of these topics include: climate change, sustainable agriculture, wildlife conservation, geopolitical tensions and antibiotic resistance. Each search query had the form “SYNONYM TOPIC”, where SYNONYM denotes a synonym/type of interactive (geo)visualization (as suggested by the LLM) and TOPIC refers to a theme (taken from the pool generated from the LLM as well). Examples of search queries are “Interactive mapping tool Roman Empire” and “GIS dashboard Vietnam War protests”. The full list of topics and search queries is available on GitHub.

Web document search: Searches with the Google Custom Search API [18] were done to retrieve urls that have a higher chance of containing an IV or IGV. The retrieved urls were saved in a MongoDB database.

Web document labelling: A Python script was created to facilitate the annotation. It launches an interactive command line that automatically takes an unlabelled webpage from the database, opens it in the browser and asks the user to provide a label. Irrelevant webpages can be also deleted from the database through the interactive command line.

The labelling of the webpages faced a few challenges. For instance, some webpages took several minutes to load and show their content, which impedes effectiveness when classifying thousands of items. Also, some pages could not be opened and were therefore unusable. These webpages were deleted from the database. Another challenge was the low recall in the early stages. After running the first set of search queries and labelling 171 items, the percentages of classifications were only around 5.8% (IV) and 2.9% (IGV) respectively. This is an improvement compared to the 0.05% reported in [3], but still not high enough for scalable dataset generation. As mentioned above, the first set of queries followed the template “SYNONYM TOPIC”. Initially, the queries were slightly verbose in the hope that these would lead to a better matching of the entities of interest, e.g. “Interactive geovisualizations Satellite technology for Earth observation”, “Map-based data exploration The Great Wall of China construction” (see the full list on GitHub). Many webpages returned after this first set of queries contained long scientific texts, notably in the form of PDF documents. In light of these initial results, the approach was changed towards more simplified search queries. Both SYNONYM and TOPIC were made more concise, e.g. “interactive map weather” and “dynamic map air pollution”. After these changes regarding the search queries, the percentage of IV classifications went up a bit to 10%. For further improvements, the data collection approach evolved once more to focus on dashboards. Dashboards used include Carto, Ceros (ceros.com), Esri (arcgis.com/apps/dashboards), Highcharts (highcharts.com/demo), Infogram (infogram.com), Plotly (plotly.com) and Tableau (tableau.com). It should be noted

■ **Table 1** Descriptive information about Geovicia: #code and #embed signal the availability of the original HTML code and their embeddings values; #featureinformation denotes semi-structured information (extracted post-harvesting) available in the dataset.

	#count	#avglen	#sdlen	#minlen	#maxlen	#code	#embed	#featureinformation
NoIV	1153	224094.8	329201.2	52	4711499	Yes	Yes	url, content, description, external links, external scripts, div_ids, class_ids
IV	476	158247.7	169970.4	1186	1323808			
IGV	910	111248.5	249832.6	52	2906885			
All	2539	171305	282034	52	4711499			

that the webpages were not solely collected from these dashboards. A portion of the dataset stems from the search results and pages linked to them. Indeed, it was often the case that a webpage contained links to other webpages with IVs or IGVs. When this was noticed while labelling the webpage, these webpages were added to the database as well.

Table 1 shows some information about the resultant dataset, which is available in two formats for reuse: CSV (Comma-Separated Value) and JSON (JavaScript Object Notation).

4 Automated Classification

As discussed in previous work [9], the automated classification of web-based documents involves two steps: webpage representation (i.e. transforming the webpage into a feature vector) and webpage classification (where machine learning models are trained/used to learn the classification function for a set of features). This section briefly presents the two steps.

4.1 Representation

Two approaches were considered to extract features from the websites, namely a feature-engineering approach and an embedding-based approach.

- **Feature-engineering approach:** The gist of the feature-engineering approach is the presence or absence of selected keywords in some portions of the web document, notably: content, description, external_links, external_scripts, div_ids and div_classes. Following Hüffer et al. [21] four types of keywords were considered: names of frameworks (e.g. highcharts, d3, leaflet), IDs of HTML elements (e.g. apexcharts, map, globe), classes of HTML elements (e.g. tableau, esri-map, mapboxgl) and sentences (e.g. interactive, geovisualization, Datenvisualisierung). The full list of keywords considered is extended from [21] and is available on GitHub. The presence/absence of these keywords is encoded using one-hot encoding, leading to a sparse vector with 74 entries. The three classes of target entities (IGV, IV and noIV) are encoded using label encoding (and more precisely the LabelBinarizer from the scikit-learn library).
- **Embedding-based approach:** Text embeddings encode text into dense vectors that capture the meaning and are useful for measuring the relatedness of text snippets. While the feature-engineering approach generates a small, transparent set of features for training machine learning models (see above), the features generated by text embedding models are opaque, as they are produced automatically. We considered both open-source and proprietary large language models for generating the embeddings. The Massive Text Embedding Benchmark [29] (MTEB) guided the selection of the open-source model. Our goal was to identify the optimal trade-off between model performance and model context length. With these aspects in mind, the model stella 1.5b (with 1024 dimensions)

was chosen². It has a memory footprint of approximately 6 GB, ranks under the first 10 models concerning classification as a task and has a token limit of 131,072 (which is the second highest of all models). BERT and GPT2 used in previous work [22] for geometry and spatial relation representations have a much lower token limits (512 and 1024 tokens respectively) and hence were not considered in this work. The same goes for recent text embedding models by OpenAI, which have a context length of about 8200 tokens [33]. About one-third of the webpages have more characters than the context length of the stella model. Hence, to assess the sensitivity of the results to context length, we report the classification results for two settings: (1) all web documents (referred to as Embedding-based I), and (2) web documents shorter than the token limit (referred to as Embedding-based II).

4.2 Classification and experimental setup

We considered five models from different families of classification algorithms: k-nearest neighbors (kNN; instance-based learning [39]), support vector machine (SVM) [39], Naive Bayes (Bayesian Network [25]), random forest (ensemble) and multi-layer perceptron (neural network [25]).

kNN: The value of k was determined using a grid search on the training set. The best parameters obtained were: k = 3, weight = uniform (feature-engineering); and k=5, weights=distance (embedding-based).

SVM: We compared the performance of the linear and the radial basis function (rbf) kernels. The rbf kernel led to no or only very minimal improvements so that the linear SVM was selected due to its simpler kernel function and its faster training time (Occam’s razor principle).

Naive Bayes: We compared a Gaussian model and a Bernoulli model. Based on the results, we selected the Bernoulli model for the feature-engineering approach and the Gaussian model for the embedding-based approach. This is also in line with theoretical considerations: The Bernoulli model relies on binary occurrence information whereas the Gaussian model assumes that values of features are normally distributed [40].

Random Forest: The best parameters obtained using grid search were: n_trees = 200 (feature-engineering) and n_trees = 400 (embedding-based).

Multi-layer Perceptron: A grid search was used to identify the best-performing architecture. The outcome was an architecture with hidden layer sizes of (36, 18, 9) for the feature-engineering approach and a shallow network with a single hidden layer with 512 neurons for the embedding-based approach.

We used the F1 score with macro averaging for decision-making in all cases because we have an imbalanced dataset. The grid search for hyperparameter fine-tuning was done using a 10-fold cross-validation. Model comparison for selection was done using 10-fold cross-validation as well. Besides, we tested two classification strategies: multiclass (IGV, IV, noIV) and binary (IGV vs noIGV), as we are primarily interested in the automated classification of web-based geovisualizations. We also assess the impact of balancing and the representation strategy (feature-engineering vs embedding-based) on performance. At last, we explore the sensitivity of the results to the threshold of the context length of the LLM-generated embeddings. We used a 80/20 % train and test data split in the experiments.

² https://huggingface.co/dunzhang/stella_en_1.5B_v5. Though the model is accessible in multiple dimensions, 1024 provided a good compromise between size and performance as of December 2024.

Tables 2 and 3 present the results for multiclass classification and binary classification respectively. The confusion matrices for the models are available as supplementary material at <https://doi.org/10.6084/m9.figshare.28238885>. To compare our results to the state-of-the-art, we include the results from a heuristic-based (i.e. rule-based) approach from [21], which was suggested for multiclass classification. The values obtained were 49% (accuracy), 54% (precision), 47% (recall), and 42% (F1 score). Finally, we used permutation feature importance, introduced originally in [6], to investigate the contributions of each feature to the overall classification accuracy in the case of the feature-engineering approach. The tests were done for the random forest and the multi-layer perceptron models, and the results are available in the supplementary material as well.

■ **Table 2** Results of the multiclass classification (IGV vs IV vs noIV). Best values are in **bold**. Embedding-based I = all documents; Embedding-based II = documents fitting Stella’s context length.

			Accuracy	Precision	Recall	F1	ROC-AUC
Feature-engineering	Imbalanced	knn	62%	71%	55%	56%	0.66
		svm	62%	71%	55%	55%	0.66
		nb	57%	66%	53%	55%	0.65
		rf	62%	69%	55%	56%	0.66
	Balanced	mlp	61%	69%	55%	56%	0.66
		knn	35%	68%	35%	46%	0.63
		svm	29%	67%	32%	43%	0.62
		nb	31%	75%	31%	41%	0.63
		rf	36%	71%	36%	47%	0.64
		mlp	34%	73%	35%	46%	0.64
Embedding-based I	Imbalanced	knn	69%	70%	71%	70%	0.78
		svm	54%	73%	53%	61%	0.71
		nb	29%	52%	85%	64%	0.73
		rf	67%	76%	65%	69%	0.77
	Balanced	mlp	62%	76%	59%	66%	0.74
		knn	70%	70%	71%	69%	0.78
		svm	69%	70%	70%	70%	0.78
		nb	40%	63%	88%	72%	0.79
		rf	67%	74%	67%	69%	0.78
		mlp	71%	73%	75%	74%	0.81
Embedding-based II	Imbalanced	knn	65%	65%	72%	67%	0.77
		svm	67%	67%	74%	69%	0.78
		nb	39%	56%	75%	64%	0.71
		rf	63%	65%	69%	66%	0.76
	Balanced	mlp	69%	70%	74%	71%	0.80
		knn	63%	65%	65%	64%	0.74
		svm	62%	62%	66%	63%	0.73
		nb	30%	56%	83%	66%	0.74
		rf	62%	67%	63%	64%	0.74
		mlp	65%	64%	67%	65%	0.75

4.3 Discussion

We now discuss the different effects assessed in the work: effect of the representation strategy, of the classification model, of the classification strategy, of balancing and of context length.

- Effect of the representation strategy: In nearly all instances, the embedding-based performances were higher than those obtained using the feature-engineering approach (F1 and ROC-AUC scores). This suggests that the embeddings were likely better at condensing

■ **Table 3** Results of the binary classification (IGV vs noIGV). Best values are in **bold**. Embedding-based I = all documents; Embedding-based II = documents fitting Stella’s context length.

			Accuracy	Precision	Recall	F1	ROC-AUC
Feature-engineering	Imbalanced	knn	72%	73%	63%	63%	0.63
		svm	73%	78%	62%	61%	0.62
		nb	73%	74%	63%	63%	0.63
		rf	72%	76%	62%	61%	0.62
		mlp	73%	77%	63%	62%	0.63
	Balanced	knn	67%	69%	67%	67%	0.67
		svm	65%	69%	65%	64%	0.65
		nb	65%	69%	65%	64%	0.65
		rf	67%	70%	67%	66%	0.67
		mlp	66%	69%	66%	65%	0.66
Embedding-based I	Imbalanced	knn	77%	75%	73%	74%	0.73
		svm	77%	76%	72%	73%	0.72
		nb	68%	67%	69%	67%	0.69
		rf	79%	79%	74%	75%	0.74
		mlp	78%	77%	74%	75%	0.74
	Balanced	knn	72%	73%	72%	72%	0.72
		svm	74%	74%	74%	74%	0.74
		nb	68%	68%	68%	67%	0.68
		rf	73%	73%	73%	73%	0.73
		mlp	75%	75%	75%	75%	0.75
Embedding-based II	Imbalanced	knn	67%	67%	66%	66%	0.66
		svm	69%	69%	68%	68%	0.68
		nb	59%	59%	59%	58%	0.59
		rf	68%	69%	67%	67%	0.67
		mlp	71%	71%	71%	71%	0.71
	Balanced	knn	62%	58%	56%	55%	0.56
		svm	64%	62%	61%	61%	0.61
		nb	56%	58%	58%	56%	0.58
		rf	67%	65%	65%	65%	0.65
		mlp	67%	65%	64%	64%	0.64

relevant features to separate the different types of entities than the hand-crafted features. Also, these results remind of the “black box conundrum” [27] – model interpretability and predictive power are often competing goals for (Geo)AI models. Another aspect to mention in the comparison of the two approaches is that the embedding-based approach is more time/resource-consuming. For example, computing one single embedding takes around 20 seconds (on a laptop with an AMD Ryzen 7 7840U processor (3.30 GHz), integrated Radeon 780M Graphics, 32 GB of RAM, and 1 TB of storage, on Windows 11), which is the reason why the embeddings were pre-computed and included in the final dataset. Features from the feature-engineering approach can be computed at run-time as the feature extraction algorithm only takes a few milliseconds to run.

- Effect of the classification model: As the tables suggest, all models have comparable performance for the feature-engineering approach. The relatively low F1 scores (40%–60%) indicate the need for further research exploring “intelligent hints” [1] for the separation of the three types of entities considered. Regarding the embedding-based approach, the Naive Bayes family exhibited the strongest recall (=probability of detection) for the multiclass classification task. The MLP exhibited a good performance across all settings often having the highest or second-highest F1 score. Values obtained were in the range [66%–75%] (imbalanced dataset) and [64%–75%] (balanced dataset). As the architectures used for testing were slightly different depending on the results of the grid search, the

recurrent good performance of MLP suggests the relevance of this model family for the issue at hand and recommends it as a starting point for further work. There are more families of classifiers that were not considered in this work (e.g. discriminant analysis, bagging, decision trees, see [16]) and more kernel functions (e.g. polynomial kernels for support vector machines) that could be further explored in future work.

- Effect of the classification strategy (all vs binary): There was no notable impact of the classification strategy on the performance. SVM and Naive Bayes seem to have performed better regarding the feature-engineering approach, but slightly less so for the embedding-based approach.
- Effect of balancing on performance: The balancing led at times to improvement, and at times to deterioration in performance. The dimension does not seem to impact the results and may be dropped in subsequent studies.
- Effect of context length: As mentioned above about one-third of the web documents considered had a size greater than the context length. Details of how exactly the Stella model treats those could not be found in the model's documentation. Besides, the definition of what exactly a token is varies (e.g. characters, words, subwords). Hence an empirical assessment of the impact of the context length was done. It appears from the results that there are small drops in performance (F1 scores, ROC-AUC scores) for several models when the dataset contains web documents within the context length only (Embedding-based II). This issue deserves further investigation in future work.

Limitations. Although the dataset is 30 times bigger than the one from previous work [21], it is still relatively small compared to standard machine learning datasets and could be extended in future work. Furthermore, though the webpages were inspected thoroughly, some visualisations were challenging to find and could have been missed because 1) some webpages have the policy that interactive charts are only available on screens of a specific size (i.e. large screens), and 2) some webpages had a dense hierarchical organization and several levels of nested content, which increased the difficulty of checking every interaction possibility. At last, we mentioned in Section 3 that a portion of the dataset came from dashboards. The extent to which these dashboards bias the performance results needs a systematic assessment in future work.

5 Conclusion and Future Work

Given the increasing availability of (interactive) maps on the Web, there is a need for techniques to automate their findability. While previous work has offered techniques for the classification of static maps (e.g. figures in digital documents, SVG-based maps, shapefile-based maps), there is still a need for the automated classification of interactive maps. To address this gap, we have compiled a dataset to study the automated classification of interactive (geo)visualizations and performed a preliminary assessment of models' performances at the classification task. The results obtained show that interactive (geo)visualization classification is indeed a challenging problem for existing models and deserves more attention in future research.

Follow-up work to this article can be done along the following lines:

Dataset: The work in this article was exploratory and hence the dataset was collected and annotated manually by one researcher only. The low hit rates observed during harvesting call for further work to improve the efficiency of the harvesting workflow. Besides, previous work [21] suggested that a crowd-sourcing approach to collect interactive geovisualization

annotations could be workable, but a large-scale dataset is still lacking. Hence, looking into crowd-sourcing-based approaches for the annotation task is an important direction for further work. The challenge here lies in simultaneously maintaining systematicity during collection, diversity of visualization types and themes, quality of the annotations, as well as producing more fine-grained annotations (e.g. the annotation should state not only if there is a visualization, but how many there are and where these are located in the web document if appropriate).

Representation and Classification: Regarding the feature-engineering approach, we only looked into content information while engineering the features. Previous work [2, 23] examined URL-based approaches to web-page classification and this could be investigated also for interactive (geo)visualization classification. Furthermore, combining link and content information is popular during classification [34] and could be considered in future work as well. For instance, interactive maps about attractions in cities have a higher likelihood to link to/be linked from tourist webpages; interactive maps covering events as they unfold (e.g. war, earthquake, election results) have a higher likelihood of being linked from news webpages; interactive geovisualizations in web-based notebooks such as Observable (observablehq.com) have a higher likelihood of linking to/being linked from other notebooks. This graph-based modelling of the interactive (geo)visualization is intriguing and worth additional exploration in future work, along with appropriate (graph neural network or end-to-end) architectures to boost the automated detection of interactive maps and geovisualizations on the Web.

References

- 1 Yaser S Abu-Mostafa. Machines that learn from hints. *Scientific American*, 272(4):64–69, 1995.
- 2 Mohammed Al-Maamari, Mahmoud Istaiti, Saber Zerhouli, Michael Dinzinger, Michael Granitzer, and Jelena Mitrović. A comprehensive dataset for webpage classification. In *Open Search Symposium 2023 (OSSYM2023)*, Geneva, Switzerland, 2023. Zenodo. doi:10.5281/zenodo.10594210.
- 3 Leilani Battle, Peitong Duan, Zachery Miranda, Dana Mukusheva, Remco Chang, and Michael Stonebraker. Beagle: automated extraction and interpretation of visualizations from the Web. In Regan L Mandryk, Mark Hancock, Mark Perry, and Anna L Cox, editors, *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI 2018)*, page 594, Montreal, Quebec, Canada, 2018. ACM. doi:10.1145/3173574.3174168.
- 4 Lars Bernard, Stephan Mäs, Matthias Müller, Christin Henzen, and Johannes Brauner. Scientific geodata infrastructures: challenges, approaches and directions. *International Journal of Digital Earth*, 7(7):613–633, August 2014. doi:10.1080/17538947.2013.781244.
- 5 Christopher Bone, Alan Ager, Ken Bunzel, and Lauren Tierney. A geospatial search engine for discovering multi-format geospatial data across the web. *International Journal of Digital Earth*, 9(1):47–62, January 2016. doi:10.1080/17538947.2014.966164.
- 6 Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001. doi:10.1023/A:1010933404324.
- 7 Ebubekir Buber and Banu Diri. Web page classification using RNN. *Procedia Computer Science*, 154:62–72, 2019. doi:10.1016/j.procs.2019.06.011.
- 8 Jamie Callan. Distributed information retrieval. In W. Bruce Croft, editor, *Advances in Information Retrieval: Recent Research from the Center for Intelligent Information Retrieval*, pages 127–150. Springer, 2002.
- 9 Ben Choi and Zhongmei Yao. Web page classification. In Wesley Chu and Tsau Young Lin, editors, *Foundations and Advances in Data Mining*, pages 221–274. Springer, 2005. doi:10.1007/11362197_9.

- 10 Fabio Crestani and Ilya Markov. Distributed information retrieval and applications. In Pavel Serdyukov, Pavel Braslavski, Sergei O. Kuznetsov, Jaap Kamps, Stefan M. Rüger, Eugene Agichtein, Ilya Segalovich, and Emine Yilmaz, editors, *Advances in Information Retrieval - 35th European Conference on IR Research (ECIR 2013)*, pages 865–868, Moscow, Russia, 2013. Springer. doi:10.1007/978-3-642-36973-5_104.
- 11 Auriol Degbelo. FAIR geovisualizations: definitions, challenges, and the road ahead. *International Journal of Geographical Information Science*, 36(6):1059–1099, June 2022. doi:10.1080/13658816.2021.1983579.
- 12 Auriol Degbelo, Benno Schmidt, Johnni Vuong, Christin Henzen, Franziska Zander, Sarah Lechler, and Bernadette Lier. Search user interaction in multi-theme map-based Applications: A preliminary assessment. In *MuC '24: Proceedings of Mensch und Computer 2024*, pages 640–645, Karlsruhe, Germany, 2024. ACM. doi:10.1145/3670653.3677474.
- 13 Laura Diaz, Albert Remke, Tomi Kauppinen, Auriol Degbelo, Theodor Foerster, Christoph Stasch, Matthes Rieke, Bastian Schaeffer, Bastian Baranski, Arne Bröring, and Andreas Wytzisk. Future SDI - Impulses from Geoinformatics research and IT trends. *International Journal of Spatial Data Infrastructures Research*, 7:378–410, 2012. doi:10.2902/1725-0463.2012.07.art18.
- 14 Evanthia Dimara and Charles Perin. What is interaction for data visualization? *IEEE Transactions on Visualization and Computer Graphics*, 26(1):119–129, January 2020. doi:10.1109/TVCG.2019.2934283.
- 15 Esri. Galerie / ArcGIS Online, 2025. Accessed: January 2025. URL: <https://www.arcgis.com/home/gallery.html>.
- 16 Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15(90):3133–3181, 2014. doi:10.5555/2627435.2697065.
- 17 Aman Goel, Matthew Michelson, and Craig A. Knoblock. Harvesting maps on the web. *International Journal on Document Analysis and Recognition (IJDAR)*, 14(4):349–372, December 2011. doi:10.1007/s10032-010-0136-2.
- 18 Google. Custom search JSON API, 2025. Accessed: January 2025. URL: <https://developers.google.com/custom-search/v1/overview>.
- 19 Amit Gupta and Rajesh Bhatia. Ensemble approach for web page classification. *Multimedia Tools and Applications*, 80(16):25219–25240, July 2021. doi:10.1007/s11042-021-10891-3.
- 20 Phil Hüffer. phuef/ma. Software (visited on 2025-07-28). URL: <https://github.com/phuef/ma/>, doi:10.4230/artifacts.24210.
- 21 Phil Hüffer, Auriol Degbelo, and Eftychia Koukouraki. Designing search engines for interactive web-based geovisualizations. In *Proceedings of the 26th AGILE Conference on Geographic Information Science (AGILE 2023)*, volume 4, page 27, Delft, The Netherlands, 2023. doi:10.5194/agile-giss-4-27-2023.
- 22 Yuhan Ji and Song Gao. Evaluating the effectiveness of large language models in representing textual descriptions of geometry and spatial relations (short paper). In Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise, editors, *12th International Conference on Geographic Information Science (GIScience 2023)*, volume 277 of *LIPICs*, pages 43:1–43:6, Leeds, United Kingdom, 2023. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. doi:10.4230/LIPICS.GISCIENCE.2023.43.
- 23 Min-Yen Kan and Hoang Oanh Nguyen Thi. Fast webpage classification using URL features. In Otthein Herzog, Hans-Jörg Schek, Norbert Fuhr, Abdur Chowdhury, and Wilfried Teiken, editors, *CIKM'05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 325–326, Bremen, Germany, 2005. ACM. doi:10.1145/1099554.1099649.
- 24 Wallace Koehler. An analysis of web page and web site constancy and permanence. *Journal of the American Society for Information Science*, 50(2):162–180, 1999. doi:10.1002/(SICI)1097-4571(1999)50:2<162::AID-ASIT>3.0.CO;2-B.

- 25 S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas. Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, 26(3):159–190, November 2006. doi:10.1007/s10462-007-9052-3.
- 26 Pei-Chun Lai and Auriol Degbelo. A comparative study of typing and speech for map metadata creation. In Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras, editors, *Proceedings of the 24th AGILE Conference on Geographic Information Science (AGILE 2021)*, pages 1–12, June 2021. doi:10.5194/agile-giss-2-7-2021.
- 27 Wenwen Li, Samantha Arundel, Song Gao, Michael Goodchild, Yingjie Hu, Shaowen Wang, and Alexander Zipf. GeoAI for science and the science of GeoAI. *Journal of Spatial Information Science*, 29:1–17, September 2024. doi:10.5311/JOSIS.2024.29.349.
- 28 Peter Morville and Jeffery Callender. *Search patterns: design for discovery*. O’Reilly Media, Inc, 2010.
- 29 Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. *CoRR*, 2023. doi:10.48550/arXiv.2210.07316.
- 30 Observable. Maps / Observable, 2025. Accessed: January 2025. URL: <https://observablehq.com/collection/@observablehq/maps>.
- 31 Aytuğ Onan. Classifier and feature set ensembles for web page classification. *Journal of Information Science*, 42(2):150–165, April 2016. doi:10.1177/0165551515591724.
- 32 OpenAI. Chatgpt, 2025. Accessed: February 2025. URL: <https://openai.com/chatgpt>.
- 33 OpenAI. Vector embeddings - Open AI API, 2025. Embedding Models v3. Accessed: January 2025. URL: <https://platform.openai.com/docs/guides/embeddings#embedding-models>.
- 34 Xiaoguang Qi and Brian D. Davison. Web page classification: Features and algorithms. *ACM Computing Surveys*, 41(2):1–31, February 2009. doi:10.1145/1459352.1459357.
- 35 Robert E. Roth. Interactive maps: What we know and what we need to know. *Journal of Spatial Information Science*, 6:59–115, 2013. doi:10.5311/JOSIS.2013.6.105.
- 36 Qingzhao Tan, Prasenjit Mitra, and C. Lee Giles. Effectively searching maps in web documents. In Mohand Boughanem, Catherine Berrut, Josiane Mothe, and Chantal Soule-Dupuy, editors, *ECIR 2009: Advances in information retrieval*, pages 162–176, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg. doi:10.1007/978-3-642-00958-7_17.
- 37 P. Travis Thompson, Sweta Ojha, Christian D. Powell, Kelly G. Pennell, and Hunter N. B. Moseley. A proposed FAIR approach for disseminating geospatial information system maps. *Scientific Data*, 10(1):389, June 2023. doi:10.1038/s41597-023-02281-1.
- 38 Volker Walter, Fen Luo, and Dieter Fritsch. Automatic map retrieval and map interpretation in the internet. In Sabine Timpf and Patrick Laube, editors, *Advances in Spatial Data Handling: Geospatial Dynamics, Geosimulation and Exploratory Visualization*, pages 209–221. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. doi:10.1007/978-3-642-32316-4_14.
- 39 Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, and Dan Steinberg. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37, January 2008. doi:10.1007/s10115-007-0114-2.
- 40 Shuo Xu. Bayesian Naïve Bayes classifiers to text classification. *Journal of Information Science*, 44(1):48–59, February 2018. doi:10.1177/0165551516677946.
- 41 Selma Ayşe Özel. A Web page classification system based on a genetic algorithm using tagged-terms as features. *Expert Systems with Applications*, 38(4):3407–3415, April 2011. doi:10.1016/j.eswa.2010.08.126.

Georeferencing Historical Maps at Scale

Rere-No-A-Rangi Pope 

Victoria University of Wellington, Aotearoa, New Zealand

Marcus Frean 

Victoria University of Wellington, Aotearoa, New Zealand

Abstract

This paper presents a novel approach to automatically georeferencing historical maps using an algorithm based on salient line intersections. Our algorithm addresses the challenges inherent in linking historical map images to contemporary cadastral data, particularly those due to temporal discrepancies, cartographic distortions, and map image noise. By extracting and comparing angular relationships between cadastral features, termed monads and dyads, we establish a robust method for performing record linkage by identifying corresponding spatial patterns across disparate datasets. We employ a Bayesian framework to quantify the likelihood of dyad matches corrupted by measurement noise. The algorithm's performance was evaluated by selecting a map image and finding putative angle correspondences from the entirety of Aotearoa New Zealand. Even when restricted to a single dyad match, >99% of candidate regions can be successfully filtered out. We discuss the implications and limitations, and suggest strategies for further enhancing the algorithm's robustness and efficiency. Our work is motivated by previous work in the areas of critical GIS, critical cartography and spatial justice and seeks to contribute to the areas of Spatial Data Science, Historical GIS and GIScience.

2012 ACM Subject Classification Applied computing → Mathematics and statistics; Applied computing → Earth and atmospheric sciences; Applied computing → Arts and humanities

Keywords and phrases Historical GIS, Georeferencing, Record Linkage, Spatial Data Justice

Digital Object Identifier 10.4230/LIPIcs.GIScience.2025.11

Acknowledgements Kei ngā mātāpuputu, ko David te puna mōhio, ko Sydney Shep te kanohi hōmiromiro, ko Rhys Owen te reo pono, kei ngā hoa, kei ngā whānau, tēnā koutou katoa.

1 Context

Introduction

Historical maps are a cartographic record of where a place was and perhaps still is. They offer a window into the past that can support indigenous communities to reconnect with their histories, language and places. Specifically, historical cadastral maps provide a record of the evolution of land interests during the colonisation of Aotearoa New Zealand in the 19th and 20th centuries. Given that georeferencing is typically a manual and time-intensive task, these rich resources are often inaccessible to most except for those who are familiar with cartography or the local histories of the places that were mapped. By labelling points in the map and calculating their real world coordinates, georeferencing enables a broader level of accessibility to the map image and the histories embedded therein. We propose an algorithm Koki Tauriterite (translated simply as determining angle equality) that leverages the ease of detecting intersecting lines in the map image and digital cadaster to perform robust record linkage between the two sources for the purposes of large scale automatic georeferencing of historical maps in Aotearoa New Zealand.



© Rere-No-A-Rangi Pope and Marcus Frean;
licensed under Creative Commons License CC-BY 4.0

13th International Conference on Geographic Information Science (GIScience 2025).

Editors: Katarzyna Sila-Nowicka, Antoni Moore, David O'Sullivan, Benjamin Adams, and Mark Gahegan;
Article No. 11; pp. 11:1–11:11



Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

De-colonial inspirations

For New Zealand Māori, historical maps allow one to locate traditional places of food gathering, villages and burial grounds that are no longer visible in today's landscape due to successive generations of *te muru me te raupatu* (land confiscation and dispossession), where historical and contextual knowledge of said places generally exists in the minds of a few [4]. Therefore historical maps enable one to do what most cannot; locate these cultural sights of significance in space. Attempts to revitalise and make accessible this knowledge are visible around the country and the utilisation of computational tools are aiding in that process of revival, for example Ngāi Tahu's community generated atlas tool *Kā Huru Manu* [11].

The processes that lead to the creation of the map in the Aotearoa New Zealand context represent a traumatic colonial history for Māori communities [4]. Power structures in the colony which sought to benefit the colonising power were perpetuated while simultaneously assimilating indigenous relationality to place via the creation of the map [12]. The role that maps played in the dispossession of Māori and their resources through the Native Land Court and other various mechanisms of the state [6] highlights a contradiction evident with the elevation of the map as an important archival record of Māori language and geographic history [5], particularly (for one of us) as a descendant of Parihaka ploughmen who were imprisoned without trial and forced into hard labour for ploughing their own confiscated land [14].

However, those same symbols of paternalistic control can also play an important role in the retrieval, preservation and eventual dissemination of Māori relationality and re-connection with place, and subsequently aid in helping to make visible that rich history to the communities from which those places belong and vice versa. Therefore, the rematriation [10] of historical maps as records of place, via the provision of wider, open access to them for indigenous communities represents a form of spatial data justice [15] that inspires and motivates this work.

2 Method

Challenges

The task amounts to record linkage between these two very different sources of data. Source 1 is a digitised image of a map – those of interest are not currently georeferenced. The state of the map and quality of digitisation is variable and, in addition, neither overall orientation nor scale can be assumed for the map since there are often scarce metadata records to accompany the map. This necessitates an approach that is therefore scale and rotation invariant. Source 2 is essentially a large list of polygons and associated geospatial information, henceforth “the cadaster”. The task of georeferencing a map amounts to finding plausible corresponding locations for an individual map, within the (large) cadaster. However the two sources use completely different representations in their raw data (images, and polygons respectively). Contrast this with astrometry for example, where both image and stellar catalogue naturally result in lists of 2d vectors which are able to be compared [13]. An effective solution will involve making a defensible correspondence between features identifiable in both map images and the cadaster.

Record linkage of cadastral data and historical map images presents a significant challenge due to inherent discrepancies between analogue and digital-borne data [16]. While both sources purportedly depict the same geographic area, they capture it at different points in time, resulting in variations in geometry present in both the survey record and the map

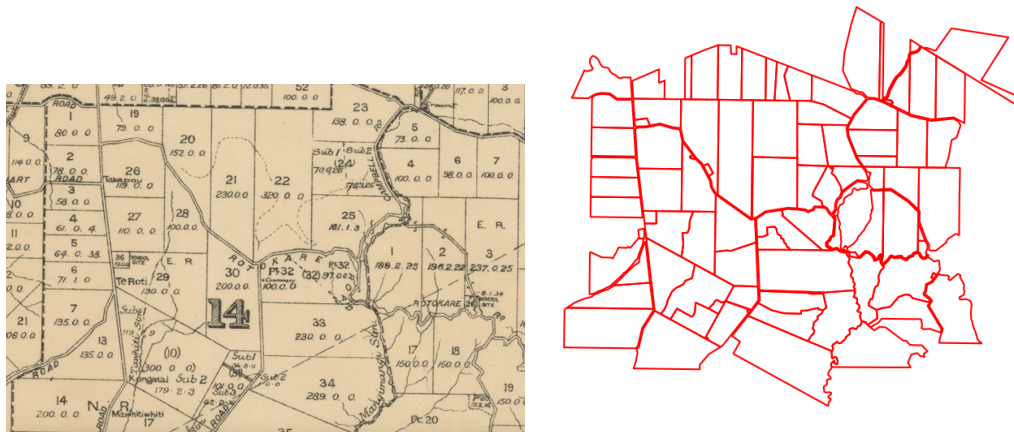


Figure 1 Map image and cadaster of the same location (NZMS13 1899, Taranaki region of New Zealand, and the current parcels retrieved from LINZ 2025 respectively).

image. Temporal changes in land use as well as cartographic conventions complicate the process of establishing accurate correspondences between the two sources of data [17]. In Aotearoa New Zealand, the cadaster is a chronological record of land title administration that has been recorded, digitised and made public by the government agency Toitū Te Whenua Land Information New Zealand (LINZ), a process which began in the mid-1980s [19]. This is the primary dataset that contains millions of polygons that represent all current parcels in Aotearoa New Zealand [18].

Herein lies another major challenge in conducting historical record linkage. While the current survey record reflects the present-day cadastral landscape, the historical map offers a snapshot of a past configuration of land ownership. Consequently, the LINZ Parcels Dataset contains numerous parcel records that are absent from the historical map due to subsequent partitions of new parcels. Therefore, the historical map may depict parcels that no longer exist in the contemporary LINZ Parcels Dataset. These disparities necessitate a robust record linkage approach accounting for the inevitable attrition of records across the temporal divide. Record linkage in this case being the process of matching and combining records from multiple sources into a single space and then subsequently assessing whether or not they are in fact describing the same object.

Cadastral features such as parcel boundary intersections are persistent to changes over time. Parcels of land are more often partitioned than they are amalgamated as a result of accelerated urban development in Aotearoa New Zealand during the 20th century [1]. Although the polygonal shape that is created by the parcel is generally orthogonally segmented over time, the original shape of the parcel remains identifiable in the cadaster. The original parcel may simply now have a number of lines drawn through it, for example parcel 22 in 1 is partitioned in the cadaster. These straight lines that depict parcel boundaries and the adjacent parcels they intersect with will be the basis for our analysis since there is no complete *ground truth* yet developed that could represent all land parcels from a particular point in time when a given map was created. Therefore a common representation space (defined below) is required to be developed upon which comparison can be performed between the two sources.

Our proposed algorithm makes use of the fact that *straight lines are particularly salient* in both sources (respectively via the Hough Transform [7, 3] for images and simple geometric processing for the cadaster). This supports ready identification of *line intersections* in either source, and characterisation by their internal angles. Agreement between sets of internal angles is useful but not diagnostic: it does not provide sufficient specificity on its own to match maps to locations. However any *pair* of intersections (which we call a “dyad” in what follows) furnishes another pair of angles defining their relative orientation, which provides a much more finely grained basis on which to argue for a match.

A “distance” that is low for good matches is simply the sum of squared differences in these two sets of angles, Θ from the map and Φ from a potential site in the cadaster:

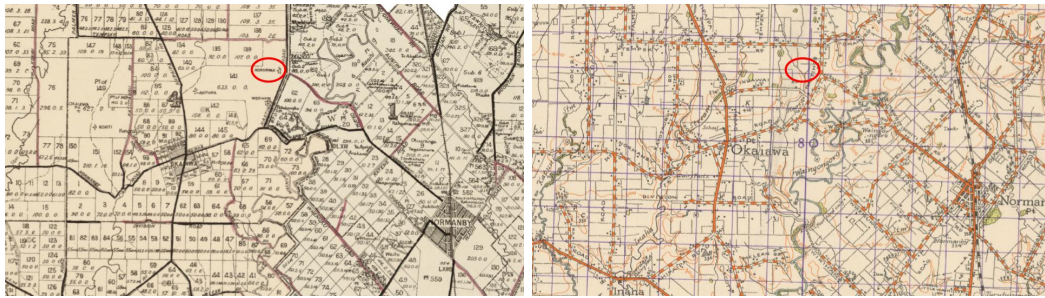
$$d(\Theta, \Phi) = \frac{1}{2\sigma_{\text{noise}}^2} \|\Phi - \Theta\|^2 \quad (1)$$

Although intuitively appealing, Equation 1 requires some justification. In Section 3, we derive it from considering a Bayes factor for the probability of a match being correct, given the respective angles Θ and Φ , and taking careful account of their various dependencies.

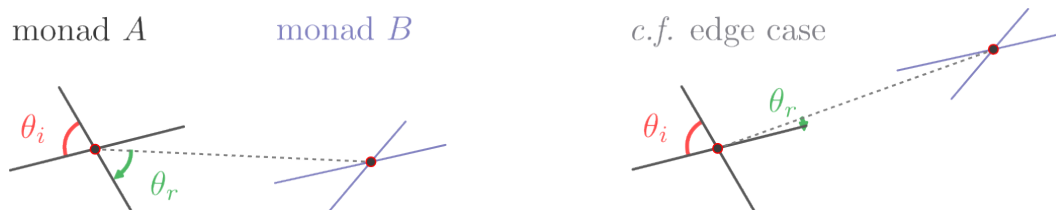
Related work

In our pursuit of an effective method for georeferencing historical map images using data from the cadastral record, we explored several approaches before developing our current line intersection-based algorithm. These preliminary investigations, while ultimately not preferred, did highlight the challenges inherent in this domain. Use of image segmentation and shape recognition techniques to match full parcel polygons is one such option, but our experience has been that these methods are not robust enough to the variability inherent to historical maps, as well as the fundamental differences in spatial object representation between the cartographer’s approach and modern GIS practices that form the current cadastral record. The discrepancies in boundary delineation and feature abstraction between these two distinct technologies rendered direct shape-matching approaches unreliable. Subsequently, we explored point-based methods, employing point-finding algorithms coupled with triangulation techniques such as the Delaunay triangulation [8]. This approach, while theoretically appealing, was also found to be unsuitable, largely due to the noise present in the map images. The arbitrary addition and removal of points caused by image artifacts and inconsistencies in feature representation led to unstable and unreliable triangulations.

We also considered leveraging toponyms extracted through Optical Character Recognition (OCR) techniques in order to conduct direct georeferencing of the historical map. As is evident in Figure 2, this approach faced limitations due to the temporal disconnect between historical and contemporary place names. Many toponyms present on historical maps have since fallen into disuse or been replaced in the process of the colonisation of Aotearoa New Zealand [2]. Text on historical maps also tends to be heavily occluded, and/or follow the geography of the feature that it is describing, introducing even more complexity into the text extraction process [9]. Nevertheless, we found value in using the more persistent toponyms, such as major town names, street names or key geographic features as a means to estimate the general location of the map and thereby reduce the search space for our dyad matching pipeline (described below). These exploratory efforts underscore the complexity of the task at hand and the need for robust, noise-resistant methods in historical map analysis and record linkage.



■ **Figure 2** Screenshots of two map images from series NZMS13 (circa 1899) and NZMS1 (circa 1959). These authoritative maps describe the same area, each printed 60 years apart. The traditional homestead Hokorima (see NZMS13 1899, red circle) is only present on the 19th century map, as well as other visible toponyms that are either no longer present or are occluded.

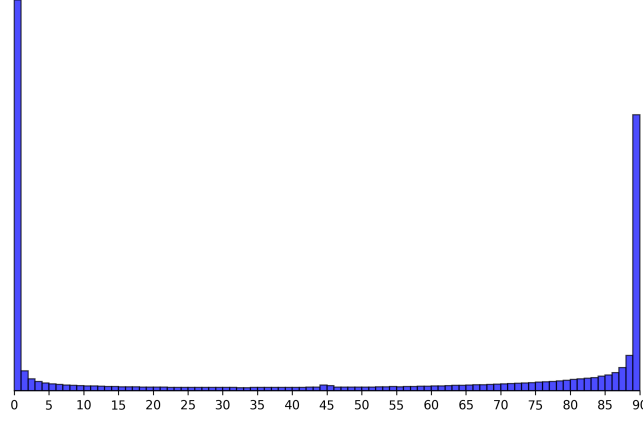


■ **Figure 3** Definition of the angles characteristic of a pair of intersecting lines. *Left:* The figure shows one dyad, and the angles for one of the monads (intersections). θ_i is the smaller of the two internal angles. Any potential second monad (here, B) creates the possibility of identifying a further “relative” angle θ_r for each monad to be the angle formed by beginning from the line connecting the two intersections and rotating clockwise to the first line encountered. Luck could easily result in θ_i values being matched within another source, but getting both θ_r correct as well becomes much less likely, for false associations. *Right:* If one dyad is displaced, both relative angles change. Here, we see that θ_r could change abruptly as monad B is shifted.

Feature extraction

In either source, we refer to a single intersection of two straight lines as a “monad”, to contrast with “dyads” (which are *pairs* of intersections) that provide the main discriminating features on which our matching algorithm is based – see Figure 3 for an example. Where the intersecting line segment does not extend past another but instead generates a “T” junction, we can still treat this as an “X” shape in effect. Each monad has a characteristic angle, defined to be the smaller of the two internal angles. This is denoted θ_i if the source is the map, and ϕ_i if it is from the cadaster. Clearly $\theta_i \leq \pi/2$ and similarly for ϕ_i .

Internal angles generated by intersecting parcel boundaries are biased toward 90° , as is evident from Figure 4. Similarly parcel boundary intersections in the cadaster result in internal angles close to 0° , due to small line segments generated by parcels that, for instance, follow the course of a winding river or are located in urbanised areas. Furthermore, the very definition of an intersecting parcel boundary in the cadaster lends itself to complexities with respect to feature extraction since it is difficult to determine whether or not a given point in the coordinate sequence of a polygon doesn’t merely sit along a straight line. For instance in most general contexts a square shaped polygon would likely produce four unique coordinates, but in the cadaster this cannot be assumed and so two parcels could intersect along an almost straight line, producing an angle very close to 0° . The very different nature of the two datasets (the historical map image being cartographic and the cadaster being



■ **Figure 4** Relative frequencies of internal angles across all monads derived from the cadaster: a large proportion of junctions are close to 0° or 90° . There are about $\sim 1.6 \times 10^6$ in the range $5^\circ < \phi_i < 85^\circ$.

generated computationally via survey measurements) leads to different complexities with respect to feature extraction since different methods for feature are required for the different datasets. Figure 4 confirms this very skewed distribution across the entire LINZ Parcels Dataset. The higher likelihood of seeing orthogonal or close to parallel lines reduces their discriminating power for the proposed algorithm, which relies on “suspicious coincidences” of angular relationships to attribute matches between dyad pairs.

A dyad is a set of two unique monads from the same source—this induces two further angles that are *relative* bearings to the line joining the monads, as illustrated in Figure 3. The range for the relative angle so defined is $\theta_r \leq \pi - \theta_i$. Collecting these angles then, a given dyad D extracted from the map image source, generated by the two monads A and B consists of $\Theta = (\theta_i^A, \theta_r^A, \theta_i^B, \theta_r^B)$, and similarly for a dyad extracted from the cadaster (which is a possible match for Θ) we have $\Phi = (\phi_i^{A'}, \phi_r^{A'}, \phi_i^{B'}, \phi_r^{B'})$. Monads and dyads are the primary features extracted from both sources. This extraction allows for comparison to be made between the two sources thereby acting as a common representational space for matching across the sources.

Koki Tauriterite: Monad and Dyad Matching Pipeline

We first pre-filter candidate matches on the basis of monad evidence alone, in order to reduce the computational burden of matching each component part of the dyad. This reduces the search space for subsequent, more computationally intensive comparisons. The pipeline can be summarised as follows:

1. **Monad matching:** For each monad in the map and cadaster, identify putative matches where internal angles differ by no more than σ degrees.
2. **Relative angle calculation:** For each plausible pairing of monads, compute the relative bearing angles θ_r and ϕ_r .
3. **Correct for edge cases:** Apply Algorithm 1 to θ_r and ϕ_r .
4. **Dyad comparison:** For each candidate dyad pair, compute a similarity score $d(\Theta, \Phi)$ based on all internal and relative angles.
5. **Best match selection:** Select the dyad match with the lowest distance score as the most likely correspondence.

The first step filters on the θ_i angles of each monad from both sources to generate a set of putative monad-monad matches where pairs meeting the filtering condition $|\theta_i - \phi_i| \leq \sigma$ are considered, with σ set to 6° . This exclusion seeks to increase computational efficiency, although we note some potential unlikely matches are overlooked. The algorithm then calculates the relative bearing θ_r or ϕ_r between pairs of monads leveraging the identified putative θ_i or ϕ_i angle matches.

Algorithm 1 is applied at this point to relative angles, in order to correct for dramatic shifts in θ_r that can occur, should the noise present in the map image (or the cadaster) completely change the θ_r angle (or ϕ_r respectively). As depicted in Figure 3, this needs to be dealt with because the θ_r angle from the connecting line to the next line segment, proceeding clockwise, can either become very small or very large given only a small amount of noise. To ameliorate this fragility we check if θ_r is very close (in either direction) to the line segments that constitute the monad, or 0. If so, we generate an edge case alternative value for θ_r , referred to as θ_{r_e} , which is then later computed during the distance step alongside θ_r . The best score resulting from the comparison of these two is treated as the true θ_r . This algorithm effectively handles the edge cases for θ_r (and similarly ϕ_r), adjusting θ_{r_e} based on its proximity to the edge of the monad's line segments, while accounting for a noise factor.

Having generated putative monad matches between the two sources, the dyad selection stage can begin. This step creates the list of relative bearing angle combinations between filtered pairs of monads that each have an inter-source putative match. The algorithm then finds potential inter-source dyad matches and can therefore start to build out potential valid inter-source tuples of dyads where each constituent monad has a putative monad match that is contained in the alternative putative dyad.

■ **Algorithm 1** Edge Case Adjustment for θ_r (and similarly for ϕ_r).

The sections mentioned are those of the monad that generate the respective internal angles (θ_i or the other, which is $\pi - \theta_i$), and b is the line that connects the two monads to generate the relative angle θ_r . If θ_r isn't close to either boundary it is returned unaffected.

procedure FINDEDGEADJUSTEDTHETAR($C, \theta_i, \theta_r, \sigma$)

$\theta_v \leftarrow \pi - \theta_i$

$\theta_{r_e} \leftarrow \theta_r$

if IntersectsWithSectionI(b) **then**

if $\theta_r \leq \sigma^2$ **then**

$\theta_{r_e} \leftarrow \pi - \theta_i - \sigma^2$

else if $\theta_i - \theta_r \leq \sigma^2$ **then**

$\theta_{r_e} \leftarrow \sigma^2$

end if

else if IntersectsWithOtherSection(b) **then**

if $\theta_r \leq \sigma^2$ **then**

$\theta_{r_e} \leftarrow \theta_i - \sigma^2$

else if $\theta_v - \theta_r \leq \sigma^2$ **then**

$\theta_{r_e} \leftarrow \sigma^2$

end if

end if

return θ_{r_e}

end procedure

3 Derivation of an appropriate distance for dyads

Here we derive Equation 1 from the standpoint of a generative model of dyads. Despite the apparent simplicity of the end result, care is needed as the relative angles are conditioned on the internal ones.

Suppose we have a single dyad from the map image source, characterised by angles Θ , and a single dyad from the cadaster, characterised by angles Φ . In addition, we have a putative association between the constituent monads delivered by the pre-processing stage: $A \sim A', B \sim B'$ where A, B are monads from a map and A', B' are from the cadaster. Taking the logarithm of the Bayes Factor (i.e., the ratio of probabilities for and against a match) yields a natural score, which quantifies the relative evidence in favour of a match on a logarithmic scale. Ignoring the prior degree of belief in a match (which just adds a constant anyway), we have a score S for a match between dyads that is the log of a likelihood ratio:

$$S(\Theta, \Phi) = \log \frac{p(\Theta, \Phi \mid \text{same})}{p(\Theta, \Phi \mid \text{diff})}$$

where “same” and “diff” refer to ground truth: the two pairs are in fact the same locations, as asserted, or are not. Using the product rule, this is

$$\begin{aligned} &= \log \underbrace{\frac{p(\Theta \mid \text{same})}{p(\Theta \mid \text{diff})}}_1 + \log \frac{p(\Phi \mid \Theta, \text{same})}{p(\Phi \mid \Theta, \text{diff})} \\ &= \log \frac{p(\phi^{A'}, \phi^D \mid \theta^A, \theta^B, \text{same})}{p(\phi^{A'}, \phi^D \mid \theta^A, \theta^B, \text{diff})} \end{aligned}$$

which factors into an A and a B term

$$= \log \frac{p(\phi^{A'} \mid \theta^A, \text{same})}{p(\phi^{A'} \mid \theta^A, \text{diff})} + \log \frac{p(\phi^{B'} \mid \theta^B, \text{same})}{p(\phi^{B'} \mid \theta^B, \text{diff})}$$

and we can unpack the two angles within each $+ BB'$ equivalents...

$$= \log \frac{p(\phi_i^{A'}, \phi_r^{A'} \mid \theta_i^A, \theta_r^A, AA' \text{ same})}{p(\phi_i^{A'}, \phi_r^{A'} \mid \theta_i^A, \theta_r^A, AA' \text{ diff})}$$

Using the product rule again,

$$= \log \frac{p(\phi_i^{A'} \mid \theta_i^A, \theta_r^A, AA' \text{ same})}{p(\phi_i^{A'} \mid \theta_i^A, \theta_r^A, AA' \text{ diff})} + \log \frac{p(\phi_r^{A'} \mid \phi_i^{A'}, \theta_i^A, \theta_r^A, AA' \text{ same})}{p(\phi_r^{A'} \mid \phi_i^{A'}, \theta_i^A, \theta_r^A, AA' \text{ diff})}$$

Consider the first term: because $\phi_i^{A'}$ is independent of θ_r^A , the latter can be dropped from the conditioning. And continuing to do this for the others as well, terms simplify to:

$$= \log \frac{p(\phi_i^{A'} \mid \theta_i^A, AA' \text{ same})}{p(\phi_i^{A'} \mid \theta_i^A, AA' \text{ diff})} + \log \frac{p(\phi_r^{A'} \mid \phi_i^{A'}, \theta_r^A, AA' \text{ same})}{p(\phi_r^{A'} \mid \phi_i^{A'}, \theta_r^A, AA' \text{ diff})}$$

We next define these four probabilities one by one, for just the AA' pair:

- $p(\phi_i^{A'} \mid \theta_i^A, AA' \text{ same})$: Here, we expect the two angles to be the same apart from measurement noise, so the probability for ϕ should be a narrow distribution centred on θ , for example $\mathcal{N}(\phi_i^{A'}; \mu = \theta_i^A, \sigma_{\text{noise}}^2)$.

- $p(\phi_i^{A'} \mid \mathbf{AA}' \text{ diff})$: uniform in the range 0 to $\pi/2$ (although in practice we ignore angles close to $\pi/2$, as noted earlier).
- $p(\phi_r^{A'} \mid \phi_i^{A'}, \mathbf{AA}' \text{ diff})$: Given that A and A' are independent, at first sight one might imagine this to be uniform in the range 0 to π , but this is not quite correct due to an internal dependence on ϕ_i :

$$p(\theta_r \mid \theta_i) = \begin{cases} \frac{2}{\pi}, & \text{if } \theta_r < \theta_i \\ \frac{1}{\pi}, & \text{if } \theta_r \text{ between } \theta_i \text{ and } \pi - \theta_i \\ 0, & \text{elsewhere.} \end{cases} \quad (2)$$

- $p(\phi_r^{A'} \mid \phi_i^{A'}, \theta_r^A, \mathbf{AA}' \text{ same})$: In most cases, similar to the comparison of internal angles for monads, we can model this as $\mathcal{N}(\phi_r^{A'}; \mu = \theta_r^A, \sigma_{\text{noise}}^2)$.

The point made in the third item above applies to the fourth as well: if $\phi_r < \phi_i$, the probability of seeing ϕ_r is twice what it would be were $\phi_r > \phi_i$. Therefore both the numerator and denominator (bullets 3 and 4) should be doubled when $\phi_r < \phi_i$. However every time this applies, it cancels perfectly, and so somewhat surprisingly we are left with just

$$S(\Theta^A, \Phi^{A'}) = \log(\mathcal{N}(\phi_i^A; \mu = \theta_i^A, \sigma_{\text{noise}}^2)) \\ + \log(\mathcal{N}(\phi_r^{A'}; \mu = \theta_r^A, \sigma_{\text{noise}}^2)) + \text{const.}$$

Logs of Gaussians yield quadratic terms. Including the BB' as well, we arrive at:

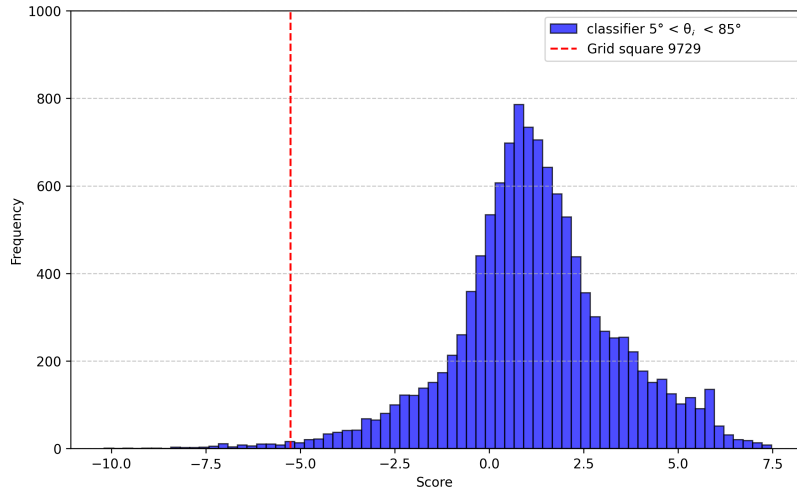
$$S(\Theta, \Phi) = -\frac{1}{2\sigma_{\text{noise}}^2} \left((\phi_i^{A'} - \theta_i^A)^2 + (\phi_r^{A'} - \theta_r^A)^2 + (\phi_i^{B'} - \theta_i^B)^2 + (\phi_r^{B'} - \theta_r^B)^2 \right)$$

Equation 1 is thus the negative log Bayes factor, which can be thought of as a distance or error. Despite the various subtleties involved, the end result is simple and intuitive: to score a match we take the sum of squared differences between (suitably defined) angles. Moreover, σ_{noise}^2 appears as a multiplier throughout, and so will not affect the relative rankings of possible matches, whatever its assumed value. In preferring matches that are close according to this measure, we are maximising the log of a likelihood ratio, under a Gaussian noise generative model for the joint distribution over angles.

4 Evaluation

To evaluate Koki Tauriterite, we tested whether it could correctly localise a single historical map image (Figure 1) within Aotearoa New Zealand. We constructed a grid of 15,870 squares covering the country, with each grid square matching the dimensions of the map image's bounding box (5.05 km \times 3.725 km). This grid served as a spatial index: for each square, we queried the LINZ Parcels Dataset and extracted the corresponding monads and dyads. This allowed nationwide coverage with a single comparison per region.

For each grid square, similarity scores were computed using the dyad matching pipeline. For the reasons given earlier, we excluded matches involving internal angles outside the range $5^\circ < \theta_i < 85^\circ$. A low distance score indicates that internal and relative angles between dyads in the grid and the map image align closely, suggesting a strong match. Despite the scale of the search – covering over 1.6 million monads in the restricted range – the algorithm correctly identified the true location of the map as one of the top-ranking candidates, scoring in the top 0.7% of all grid squares (Figure 5). This result confirms that even a single dyad match can serve as a reliable signal, enabling accurate georeferencing across tens of thousands of possible locations.



■ **Figure 5** Distribution of the best dyad match scores (the logarithm of equation 1) across all grid squares. The score of the correct region is highlighted with a red dashed line and ranks in the top 0.7% of all candidates. Note the x -axis's overall scale and offset are irrelevant as they depend on (unknowns) σ_{noise} and the prior probability of a match, respectively.

5 Conclusion

Evaluation of the algorithm's performance against the entire cadastral record of Aotearoa New Zealand highlights early success for the proposed approach. The result presented here – identifying the correct region comfortably within the top 1% of all grid squares – was achieved using only a single dyad and no additional “clues” beyond a rough estimate of scale. This suggests that the angular relationships extracted from map and cadaster are sufficiently distinctive, even at national scale, to support robust matching under favourable conditions.

Nonetheless, further improvements are needed to increase robustness and scalability in more challenging settings. In densely partitioned urban areas, the sheer number of potential dyads increases the likelihood of coincidental matches, due in part to the highly skewed angle distributions shown in Figure 4. In such saturated regions, the discriminative power of a single dyad may be insufficient to distinguish true matches from plausible distractors.

We propose two complementary strategies to address these limitations. First, the search space can be constrained by leveraging persistent toponyms extracted from the map image using OCR. These can be cross-referenced against gazetteers or other open geographic datasets, enabling the algorithm to focus only on regions plausibly represented in the map.

Second, the scoring framework can readily incorporate multiple dyads. As described in Section 3, the current algorithm assigns a score to each dyad pair based on the angular similarity of their constituent monads. Under a probabilistic interpretation, the inclusion of an additional, spatially independent dyad multiplies the strength of the match, because the likelihood of two such matches occurring by chance is the product of their individual probabilities. This compounding effect suggests multiple dyads may significantly reduce false positives and sharpen the algorithm's discriminative power. Each additional dyad imposes another independent constraint, tightening the inference and improving localisation.

We are currently investigating these extensions as part of ongoing work, as they represent a natural progression toward generalising the algorithm across a wider range of maps, including those with more noise, distortion, or limited cadastral distinctiveness.

References

- 1 Larissa Lutchman Arthur Grimes, Eyal Apatov and Anna Robinson. Eighty years of urban development in new zealand: impacts of economic and natural factors. *New Zealand Economic Papers*, 50(3):303–322, 2016. doi:10.1080/00779954.2016.1193554.
- 2 Te Aue. Davis, Tipene. O'Regan, and John. Wilson. *Nga tohu pumahara = The survey pegs of the past : understanding Maori place names*. Survey pegs of the past. The Board, Wellington, N.Z, 1990.
- 3 Richard O Duda and Peter E Hart. Use of the hough transformation to detect lines and curves in pictures. *Communications of the ACM*, 15(1):11–15, 1972. doi:10.1145/361237.361242.
- 4 Margaret Forster and Peter Meihana. Pouwhenua: Marking and storying the ancestral landscape. *Ethical Space: International Journal of Communication Ethics*, 2023(2/3), August 2023. <https://ethicalspace.pubpub.org/pub/p5xo5o1a>.
- 5 Hauiti Hakopa. *The paepae: spatial information technologies and the geography of narratives*. PhD thesis, University of Otago, 2011. URL: <https://ourarchive.otago.ac.nz/handle/10523/1801>.
- 6 Christopher Hilliard. 204the native land court: Making property in nineteenth-century new zealand. In *Native Claims: Indigenous Law against Empire, 1500–1920*. Oxford University Press, November 2011. doi:10.1093/acprof:oso/9780199794850.003.0009.
- 7 Paul VC Hough. Method and means for recognizing complex patterns, December 1962. US Patent 3,069,654.
- 8 Yasushi Ito. *Delaunay Triangulation*, pages 332–334. Springer Berlin Heidelberg, Berlin, Heidelberg, 2015. doi:10.1007/978-3-540-70529-1_314.
- 9 Jina Kim, Zekun Li, Yijun Lin, Min Namgung, Leeje Jang, and Yao-Yi Chiang. The mapKurator system: A complete pipeline for extracting and linking text from historical maps, 2023. arXiv: 2306.17059 [cs.AI]. doi:10.48550/arXiv.2306.17059.
- 10 Steven Newcomb. Healing, restoration, and rematriation, 1995. URL: <http://ili.nativeweb.org/perspect.html>.
- 11 Te Rūnanga o Ngāi Tahu. Kā huru manu, 2023. URL: <https://kahurumanu.co.nz/>.
- 12 Mark Palmer and Cadey Korson. Decolonizing world heritage maps using indigenous toponyms, stories, and interpretive attributes. *Cartographica*, 55(3):183–192, 2020. doi:10.3138/cart-2019-0014.
- 13 Jeffrey R Pier, Jeffrey A Munn, Robert B Hindsley, GS Hennessy, Stephen M Kent, Robert H Lupton, and Željko Ivezić. Astrometric calibration of the sloan digital sky survey. *The Astronomical Journal*, 125(3):1559, 2003.
- 14 Dick Scott. *Ask That Mountain: The Story of Parihaka*. Raupo Publishing (NZ) Ltd, Auckland, New Zealand, 2008.
- 15 Edward W. Soja. The city and spatial justice. *justice spatiale / spatial justice*, September 2009. [«La ville et la justice spatiale», traduction : Sophie Didier, Frédéric Dufaux]. URL: <http://www.jssj.org/>.
- 16 Ian Winchester. The linkage of historical records by man and computer: Techniques and problems. *The Journal of Interdisciplinary History*, 1(1):107–124, 1970. URL: <http://www.jstor.org/stable/202411>.
- 17 Anders Wästfelt. Ambiguous use of geographical information systems for the rectification of large-scale geometric maps. *The Cartographic Journal*, 57(3):209–220, 2020. doi:10.1080/00087041.2019.1660511.
- 18 Toitū Te Whenua Land Information New Zealand. Landonline: Parcels, 2024. data retrieved from Land Information New Zealand, <https://data.linz.govt.nz/layer/51976-landonline-parcel/>.
- 19 Toitū Te Whenua Land Information New Zealand. Historic property databases, 2025. URL: <https://www.linz.govt.nz/products-services/data/types-linz-data/property-ownership-and-boundary-data/historic-property-databases>.

Large Multi-Modal Model Cartographic Map Comprehension for Textual Locality Georeferencing

Kalana Wijegunaratna ✉ 

School of Mathematical and Computational Sciences, Massey University, Auckland, New Zealand

Kristin Stock ✉ 

School of Mathematical and Computational Sciences, Massey University, Auckland, New Zealand

Christopher B. Jones ✉ 

School of Computer Science and Informatics, Cardiff University, UK

Abstract

Millions of biological sample records collected in the last few centuries archived in natural history collections are un-georeferenced. Georeferencing complex locality descriptions associated with these collection samples is a highly labour-intensive task collection agencies struggle with. None of the existing automated methods exploit maps that are an essential tool for georeferencing complex relations. We present preliminary experiments and results of a novel method that exploits multi-modal capabilities of recent Large Multi-Modal Models (LMM). This method enables the model to visually contextualize spatial relations it reads in the locality description. We use a grid-based approach to adapt these auto-regressive models for this task in a zero-shot setting. Our experiments conducted on a small manually annotated dataset show impressive results for our approach (~1 km Average distance error) compared to uni-modal georeferencing with Large Language Models and existing georeferencing tools. The paper also discusses the findings of the experiments in light of an LMM's ability to comprehend fine-grained maps. Motivated by these results, a practical framework is proposed to integrate this method into a georeferencing workflow.

2012 ACM Subject Classification Computing methodologies → Visual inspection

Keywords and phrases Large Multi-Modal Models, Large Language Models, LLM, Georeferencing, Natural History collections

Digital Object Identifier 10.4230/LIPIcs.GIScience.2025.12

Supplementary Material Dataset: <https://doi.org/10.6084/m9.figshare.29093882.v1>

Funding This research was partly funded by the Ministry of Business Innovation and Employment Smart Ideas Fund under the BioWhere Project (grant number MAUX2104).

1 Introduction

Georeferencing is the process of relating or interpreting information to a geographic location [20, 7, 19]. *Informal* georeferencing is the association of information with a location using place names (also called toponyms) or location descriptions from ordinary human discourse. On the other hand, *formal* georeferencing refers to exact locations using formal quantitative representations such as latitude and longitude coordinates or other spatial referencing systems [20]. The task of converting an *informal* georeference to a *formal* georeference can be challenging due to reasons such as colloquial place names, outdated names, historical places, the use of vague relative spatial relations, and differences in place representations in different gazetteers (geospatial databases).

A vast amount of information is locked up in extensive collections of unstructured textual data that is yet to be systematically georeferenced. These collections include but are not limited to web pages, social media articles, academic research articles, biological collection specimen records, and memoirs. The ubiquity of georeferencing has led to numerous



© Kalana Wijegunaratna, Kristin Stock, and Christopher B. Jones;
licensed under Creative Commons License CC-BY 4.0

13th International Conference on Geographic Information Science (GIScience 2025).

Editors: Katarzyna Sila-Nowicka, Antoni Moore, David O'Sullivan, Benjamin Adams, and Mark Gahegan;

Article No. 12; pp. 12:1–12:19



Leibniz International Proceedings in Informatics

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

georeferencing techniques adopted in various application domains. For example, attempts have been made to georeference social media posts, social media images, satellite and aerial images, web documents, and collection records from natural history archives [61, 18, 44, 19, 37]. In this study, we focus on georeferencing textual locality descriptions in records of natural history specimens found in museum and herbarium archives, where it is estimated that of the order of 3 billion records are preserved [4]. It is also estimated that manual georeferencing of digital records without coordinates held globally could take over 5000 person-years [49].

A locality description is a textual description of the location at which a biological or other sample was collected. These descriptions are part of the information recorded about the specimen or sample by the collector and, for millions of pre-GPS collection records, they can be the only detailed information about the collection location. Georeferencing such locality descriptions for purposes of biodiversity studies is a considerable challenge, especially due to their sheer volume and the descriptions themselves often employing quite complex language with one or more relative spatial relations [36]. Much of the published literature on georeferencing entire sentences has focused on social media posts, with the more advanced methods using various forms of language models including transformer models [56]. Methods developed for georeferencing social media posts can also rely heavily on metadata, such as the user network. The locality descriptions with which we are concerned differ significantly from the text of social media postings in their frequent use of relative descriptions often with multiple reference named places, and where the described location is separate (offset) from that of the finer-grained place names. Several studies have focused on the development of methods to georeference such relative locality descriptions in natural history records but little progress has been made to date on the application of current deep learning methods.

Figure 1 provides an example of a locality description. Given this quite specific locality description, a human georeferencer can locate this collection location to a high degree of certainty. Manual georeferencing uses the place names in a locality description to focus on a map that covers the local area to which the description applies. Visualization of the spatial configuration of the named places is vital to a human georeferencer in identifying a point or region of space that appears to correspond to the described location [36]. However, none of the existing automated textual georeferencing methods exploit maps directly. Gazetteer lookup methods only rely primarily on locations of place names, though they can be combined with methods that compute spatial relations [18, 8]. Current deep learning based methods for georeferencing can use pre-trained language models like BERT [11] that have been pre-trained on masked language modeling and next sentence prediction. They rely on fine-tuning these pre-trained models exposing them to large numbers of example texts with their associated locations [44, 30]. Although language models can be adept at learning textual relations, being trained only on language tasks, they do not intrinsically grasp spatial dimensions. The models also do not comprehend spatial extents of the features they are working with. Furthermore, a georeferencing language model trained on one region or country can not be used to infer localities from a different region, requiring more fine-tuning and large volumes of verified data from each region. Additionally, no research appears to have been published to date on using the latest Large Language Models (LLM) for this task.

Here we present initial investigations of the potential of Large Multi-Modal models (LMM), that can support tasks combining language and vision, to assist in the georeferencing process for complex locality descriptions. With an LMM's multi-sensory skills, we experiment with a prompting approach that emulates the way that a human might georeference such

¹ <https://www.landcareresearch.co.nz/tools-and-resources/collections/allan-herbarium/>

J.K. Donald Wildlife Reserve, NE shore of L. Wairarapa - about 400m from lake

■ **Figure 1** A well defined example locality description from a collection held by the Allen Hebarium¹. Green and purple indicate place names and relative spatial indicators respectively. Here, “lake” is a coreference to Lake Wairarapa.

descriptions. The intuition in this study is to combine conventional text-based prompting with a map excerpt corresponding to the described location. This exploits the LMM’s superior language capabilities while testing its vision encoder for its map reading ability. As current state-of-the-art LMMs excel in language generation and do not perform image segmentation, we superimpose on the map a grid with labelled cells and prompt the LMM to identify the grid cell of the target location. The LMM is given the locality description, the map and the size of the grid cells. We present the results of this study comparing to an existing method, designed for interpreting locality descriptions, and other approaches to using LLMs. Motivated by these results, we design and describe a workflow that can be used to practically automate georeferencing. While the complete workflow is work-in-progress, the core georeferencing module and other elements are already in use for experiments.

Section 2 of the paper will present the related work, after which we will discuss the framework developed to use LMMs in georeferencing in Section 3. Section 4 presents the experiments, results and discussion followed by the conclusion in Section 5.

2 Related work

2.1 Georeferencing

The earliest methods for georeferencing text were based on detecting and geocoding place names in the text, which could then be used to assign one or more spatial footprints. Numerous methods for this detection and geocoding process (sometimes referred to jointly as *geoparsing*) have been developed [16, 58], and some of these have used deep learning approaches. In the case of [15], input to a convolutional neural network included the place names, context words and target name, and a vector representation of a pixel map of place name instances, that assisted the disambiguation process. Document georeferencing methods are currently dominated by language modelling approaches that treat all terms in a text document as evidence for its location [37]. The initial language models used Bayesian modelling to associate words with locations, where the locations could be grid cells [46, 62], or clusters [55], where the latter included snapping the location to the most similar already georeferenced existing document (in their case a social media posting). More recently, transformer language models have been adopted either to infer coordinates with a regression approach [44] or to classify a location as a geographic region [47], or a point of interest [30].

None of the methods above were specifically intended to deal with relative location descriptions such as commonly occur in archived natural history records. Several studies have presented rule-based approaches to georeferencing natural history specimen locality descriptions that use relative spatial relations to specify an offset relative to a reference place name. Different sorts of offset include simply distance from a named object, distance in a specified cardinal direction, and distance along a path. Typically these methods include some or all of detecting place names and spatial relational phrases, disambiguating and hence geocoding the place name, applying the offset distance, and computing some measure of uncertainty. The point radius method [61] was developed to achieve this, in which offsets

were calculated relative to a representative point of a feature while also taking account of its extent. The uncertainty of an inferred point-based georeference was expressed as a radial distance that is a function of the six factors of extent of the locality, distance precision, direction precision, unknown datum, coordinate measurement precision and map scale.

The point radius approach was refined in [18, 34], by defining several types of density based uncertainty fields, that take into account the shape of the reference object and which can be combined for complex descriptions. [53] computed distance and direction offsets, accompanied by the spatial minimality toponym disambiguation method [27], and applying a confidence measure based on matching the target record to already georeferenced records of the same survey expedition, and to the nearest location of other archived records that have the same species.

Georeferencing of descriptions of locations that use spatial relations and which were generated in a human subject experiment was described in [8]. This is one of the few examples of developing and experimenting with geospatial models of spatial relations in natural language expressions outside of the natural history domain. The approach combined models of the applicability of different sorts of relative spatial relations and required the prior existence of a place graph of the spatial relationships between places mentioned in the texts.

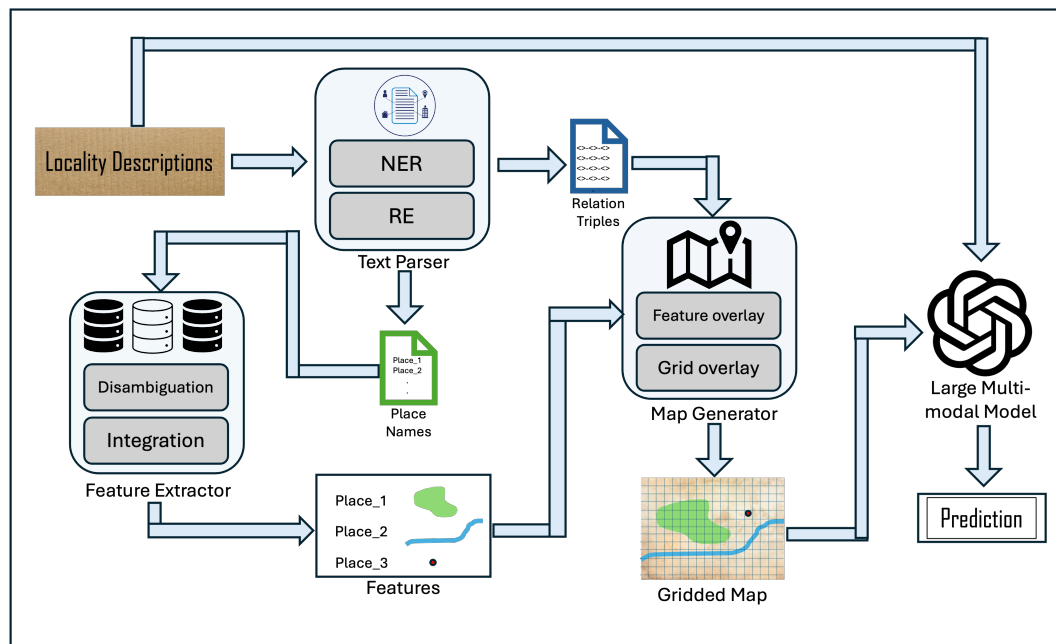
2.2 LMMs and Geospatial Use Cases

With the recent rapid development of LLMs such as GPT4 [1], Llama [52], PaLM [10], Flamingo [2], and DeepSeek’s V3 [32] and R1 models [17], adding other modalities, including vision, was seen by many as the next improvement. This led to the development of LMMs such as GPT-4Vision [40], Qwen-vl [5], PALM-E [12], Gemini-Pro Vision², Sphinx and Janus Pro [9]. However, there exist Vision-Language models that predate these LMMs such as CLIP [41], LLaVa [33] and BLIP [29] that combine the two modalities. These models have set benchmarks in various Vision-language tasks such as Visual Question Answering (VQA) [3, 24], image captioning [45, 39], visual language navigation [48] and visual reasoning [65].

LMMs have been applied in several geospatial applications. Vision capable models like GPT-4 Vision, Gemini Pro Vision, and Sphinx have been tested for tasks like map element recognition, where GPT4Vision has proved superior [63]. This study also tests GPT4Vision’s comprehension of thematic maps, point pattern, and time series analyses. GPT4Vision has also been tested in its ability to understand weather charts and make forecasts [26]. Although not using vision capabilities, LLM’s abilities to carry out spatial tasks like mapping using code and external tools like MapBox³, spatial reasoning, and describing interior locations have been tested [21, 31]. Perhaps the study closest to ours in use case is [71], although they do not use Language-Vision models. This study focuses on geolocating images. They consider maps and image embeddings as two modalities in their multi-modal fusion approach, where they use maps to build a point-cloud representation that can be fused with embeddings from images to exploit heights of buildings to better geolocate images. To the best of our knowledge no method attempts to georeference textual locality descriptions or any form of text documents with LMMs using maps as inputs. We were also unable to find any literature attempting to georeference textual documents using LLMs.

² <https://aistudio.google.com/>

³ <https://www.mapbox.com/>



■ **Figure 2** Workflow of the complete automated georeferencing process.

3 Methodology

Figure 2 presents the overall workflow of our proposed framework to utilize large multi-modal models to accurately georeference locality descriptions using gridded maps. We present a detailed description of the proposed method and the individual modules in this section.

3.1 Textual Information Parsing

As illustrated in Figure 2, the first step of the process is to extract the names of the places. Grounding named places is the most effective and simplest form of georeferencing and this is vital to our workflow. Named Entity Recognition (NER) [38] is an extensively researched problem in Natural Language Processing (NLP). Place names or locations are one of the classical semantic types that NER uses to assign labels to tokens or words [28], making most NER solutions accessible for this step of our framework. Off the shelf NER tools such as spaCy⁴, StanfordNER [14], NLTK [6], and attention [56] based pre-trained transformer models [70, 50] or modern LLM based approaches [22, 13, 66] can be leveraged for the recognition of place names. Coreference resolution [51] can be beneficial when parsing relations as illustrated in Figure 1. The extracted entities are used for Relation Extraction (RE) and finally passed to the Feature Extractor module.

The subsequent step is the extraction of spatial relations between entities. As illustrated in Figure 1, a single locality description may contain multiple relation clauses in the form of $\langle locatum, spatial\ indicator, relatum \rangle$ triples that relate a location or located object (the *locatum*) to a reference object or location (the *relatum*) with a phrase or clause denoting the spatial relationship (*spatial indicator*). It is also common in locality descriptions to have

⁴ <https://spacy.io/>

degenerate spatial relations where the locatum is not explicitly mentioned in text but is often the final location being described [23]. RE is also a thoroughly studied area. In addition to generic RE methods [70, 57, 67] used in information extraction and NLP, more geospatial relation oriented RE methods have also been developed [25, 35]. Relation triples extracted using these methods will then be passed to the Map Generator module (see Figure 2).

3.2 Geospatial Feature Extractor

Gazetteers and geospatial databases serve as fundamental resources for the grounding of place names, providing structured and authoritative spatial references. This module will be responsible for extracting relevant features from these knowledge bases, disambiguating them, and selecting the preferred representation of the place instance. While individual states often maintain authoritative gazetteers, several prominent sources provide global coverage. These include, but are not limited to OpenStreetMap⁵ (OSM), GeoNames⁶ and, Google Places API⁷. These sources can vary in their coverage of different place categories (e.g., natural features vs. artificial structures) and in the type of geometric representations they offer, ranging from point-based locations to more complex polygonal and linear footprints. The reliability and completeness of these sources can also vary as some of them are authoritative while others are community-based volunteered information. As the collection country and region are usually included in the records held by collection agencies, we are also able to exploit country-specific gazetteers, allowing us to draw from more authoritative and accurate sources. Conflating these sources provides the most comprehensive set of features for place names mentioned in a locality description.

First, we query the spatial databases with the place names returned by the previous module. The country name and region of collection can be used for disambiguation. If multiple candidates from the same region from the same source remain, a spatial clustering disambiguation is carried out (*cf* [27]). This clusters all place names mentioned and selects the candidates that form the strongest cluster, filtering out outliers. Subsequently, we are left with a single feature from each source per place name. In our conflation of sources, we prioritize features with complex geometries as this preserves information like extent and boundaries required for visual georeferencing. Preference is also given to authoritative sources. Finally, the selected features are passed on to the Map Generator module.

3.3 Map generation

For the effective application of LMMs in georeferencing, the creation of a map excerpt that is likely to contain the ground truth sample collection location is essential. As the first step of the map generation process, our map server will overlay the features returned from the Feature Extractor on a suitable basemap. Also vital to accurate georeferencing using a vision-based approach is the scale of the map. The map excerpt should be created with all essential landmarks and features necessary for an accurate georeferencing. It should also not be too coarse-grained, to avoid very large grid cells and high uncertainty. We propose the following steps to create the map excerpts:

1. In a location description with two or more named places, x, y where location x is completely contained in y , the full extent of y need not be included in the map extract. Take for example, the following locality description: ***North Island, Bay of Islands County***.

⁵ <https://www.openstreetmap.org/>

⁶ <https://www.geonames.org/>

⁷ <https://developers.google.com/maps/documentation/places/web-service>

Ca 2km north of Puketi. In this example, North Island contains Bay of Islands County and the county contains Puketi, a small locality. We avoid creating a much coarser grained map by not including the whole extent of the North Island or the Bay of Islands region and focusing on the most fine grained location (Puketi). However, the parent entity is used for disambiguation purposes when retrieving the child entity.

2. If there are two or more independent locations at the same level, the map extract must include the full extent of all such features. e.g.: *Fiordland, Mount George, south shore of lake at head of Elizabeth Burn, 2km north of peak.* In this example, both Elizabeth Burn and Mount George are included in the map excerpt. The full extent of Fiordland does not need to be included as per 1. above.
3. If the description includes an absolute distance based spatial relation, we ensure the map excerpt includes a buffered spatial extent of the relatum.
4. We ensure features are clearly visible in contrast to the base map. i.e. distinct boundaries for polygon features, clearly highlighted linear and point features.
5. We ensure legible labels for all identified and retrieved places.

Subsequently, we superimpose a labeled square grid on the map excerpt. We also record the size of the map grid cells as this is used during inference to calculate relative distances.

3.4 Multi-modal Georeferencer

The Georeferencer, essentially a Large Multi-modal Model pre-trained on both language and vision tasks, is the core of the proposed framework. This module takes as input the original locality description that is to be georeferenced along with the gridded map excerpt created by the Map Generator and attempts to predict a grid cell that is most likely to contain the location described in the locality description. Similar to LLMs, LMMs can be sensitive to the prompts used.

3.4.1 Prompt Design

We experimented with several prompts to choose the most effective prompt for this multi-modal georeferencing task.

1. Simple Zero-Shot Prompting [42]:

What grid cell/cells represent the following location description?
Location Description:

2. Automatic Chain-of-thought [68, 59]:

Based on the gridded map given, what grid cell/cells represent the following location description? Think step by step.
Location Description:

3. Logical Chain-of-Thought Prompting [69]:

Based on the gridded map given, what grid cell/cells represent the following location description?
Think step by step. Identify the locations mentioned and use the relative spatial relations mentioned in the description.
Location Description:

4. Logical Chain-of-Thought Prompting with grid size:

Based on the gridded map given, what grid cell/cells represent the following location description?
 Each grid cell is $\langle \text{grid size} \rangle \times \langle \text{grid size} \rangle$.
 Think step by step. Identify the locations mentioned. If a distance is mentioned in the description, use the grid sizes to calculate the relative distances.
 Location Description:

5. Persona [60] with Logical Chain-of-Thought Prompting with grid size:

You are a language and cartography expert. Based on the gridded map given, what grid cell/cells represent the following location description?
 Each grid cell is $\langle \text{grid size} \rangle \times \langle \text{grid size} \rangle$.
 Think step by step. Identify the locations mentioned. If a distance is mentioned in the description, use the grid sizes to calculate the relative distances.
 Location Description:

Our preliminary analysis of these prompting patterns indicated that the Logical Chain-of-thought prompt enhanced with the grid size produced the best results. We will carry out the rest of the experiments with this prompt.

The whole framework proposed in this section is highly reliant on the capability of an LMM to effectively and accurately georeference locations with the aid of a visual map. We present the experiments we conducted to gauge the potential of a multi-modal approach and the merits of diverging from traditional uni-modal text based approaches in the next section.

4 Experiments

4.1 Data

For our preliminary experiment, collection records were obtained from Global Biodiversity Information Facility⁸ (GBIF). GBIF collection records report accurate coordinates for 83% of the georeferenced records held in it [64]. Short location descriptions are more likely to contain only a single place name or a sequence of place names and no explicit spatial relations (though a comma separated sequence could represent a containment hierarchy). In the absence of descriptive spatial relations, any georeferencing method can, in the best case, only provide the coordinates of the place name mentioned (similar to a gazetteer lookup method). Therefore, the data were first filtered to collect location descriptions that were 60 characters or longer in length, allowing us to gauge the methods' performances on descriptive spatial relations. Given the vast number of collection records and collection types in GBIF, we limited the data to floral specimen collection records from New Zealand provided to GBIF by the Allen Herbarium. The place names and relations were manually annotated as the Text Parser was not implemented at the time of experiment.

For the purposes of this preliminary study, we randomly sampled 25 records to create cartographic map snippets. For this manually curated dataset, we only used OSM to identify named places that are overlaid on the standard OSM base map. For this experiment, we

⁸ <http://www.gbif.org>

manually checked the excerpts to ensure that the ground truth location was contained within the map excerpt. In our dataset of 25 examples, we observed that the ground truth location was consistently included within the map excerpt generated using the aforementioned steps, without needing any further manual intervention. However, it was observed that in examples with linear features extending over large geographic extents such as highways and rivers, the map excerpt was too coarse grained. In these cases, we manually zoomed in on the non-linear features in the description, making sure to preserve some sections of the linear feature. We will analyse the affects of this manual manipulation in Section 4.5.2.

Finally, each data item, e_i , in our dataset can be characterised as follows:

$$e_i = \{text_i, country_i, region_i, map_i, location_i, label_i, scale_i\}, \quad (1)$$

where *text* is the locality description, *country* and *region* are fields acquired from GBIF, *map* is the grid-labeled map, *location* is the ground truth point location of collection as recorded in GBIF (latitude and longitude coordinate pair), *label* is the label of the grid cell that contains the *location* and *scale* is the size of the grid cell in the map. We manually annotated *label* for each of these examples after the grid is superimposed. To the best of our knowledge, this is the first publicly available dataset⁹ for fine-grained cartographic map comprehension for LMMs.

4.2 Baselines

GeoImp [54] is perhaps the most recent georeferencing tool for biological specimen georeferencing but unfortunately it is no longer available online. The most effective methods developed for social media post georeferencing (such as Tweets) rely on the metadata and social network information and are therefore unsuitable for our task. GEOLocate [43] is an easy-to-use georeferencing system designed specifically for georeferencing natural history collection data, accessible both as a standalone software and an online service. We use this as one of our baselines. GEOLocate enables multiple predictions for each location description, but we only use its best prediction for this study. Additionally, as we are testing the performance of LMMs, we implement our own LLM baselines. All baselines compared against our LMM based generative approach are listed here:

1. **GEOLocate_{text}**: We use GEOLocate’s batch processing function over their online service. We only provide the textual description, $text_i$, to the service.
2. **GEOLocate_{text+region}**: With this baseline, in addition to the text to georeference, we provide GEOLocate the $country_i$ and $region_i$ from our dataset.
3. **ChatGPT_{text}**: Zero-shot georeferencing with OpenAI’s ChatGPT¹⁰. We use their flagship model, GPT-4o. We manually prompt it adapting a persona prompting pattern [60]:

You are a language and geography expert.
 Georeference the following location description. Answer with coordinates in decimal degrees.
 Location Description: $\{text_i\}$

⁹ <https://doi.org/10.6084/m9.figshare.29093882.v1>

¹⁰ <https://chatgpt.com/>

4. **ChatGPT_{text+region}**: This method takes a similar approach to **ChatGPT_{text}** but enriches the prompt with more context by explicitly providing it with the country and region of collection.

You are a language and geography expert.
 Georeference the following location description. Answer with coordinates in decimal degrees. The country and the district of the location are provided.
 This location is in $\{region_i\}$, $\{country_i\}$.
 Location Description: $\{text_i\}$

5. **GPT-4o_{text}**: We use the same prompt as the ChatGPT_{text} and the same underlying model (GPT-4o) but instead of using the web browser, we use the OpenAI's API. The distinction between the two methods is that ChatGPT_{text} has the capability to search the web and retrieve the coordinates of the place names and related information, whereas GPT-4o_{text}, accessed via the API, lacks this functionality.
6. **GPT-4o_{text+region}**: Prompts the GPT-4o model through OpenAI's API using the region and country enhanced prompt as seen in ChatGPT_{text+region}.

4.3 Evaluation Metrics

While distance to ground truth location from the prediction is a straight-forward measure of error for methods that predict coordinates, the measurement of error is slightly more complicated for comparing grid cells with coordinates. We implement three Euclidean distance metrics to calculate the distance error given the correct grid cell label, $label_i$, a predicted grid cell label, $pred_i$, and $scale_i$:

$$centroid - distance = \sqrt{|x_2 - x_1|^2 + |y_2 - y_1|^2} \times scale_i, \quad (2)$$

$$max - distance = \sqrt{(|x_2 - x_1| + 1)^2 + (|y_2 - y_1| + 1)^2} \times scale_i, \quad (3)$$

$$min - distance = \sqrt{\min(|x_2 - x_1| - 1, |x_2 - x_1|)^2 + \min(|y_2 - y_1| - 1, |y_2 - y_1|)^2} \times scale_i, \quad (4)$$

where (x_1, y_1) and (x_2, y_2) are two dimensional indices of the grid cells of $label_i$ and $pred_i$, respectively. Each grid cell is a unit square such that $(x_1, y_1), (x_2, y_2) \in \mathbb{N}^+ \times \mathbb{N}^+$. The *centroid - distance* calculates the Euclidean distance between the two grid cell centroids, where one centroid is considered the ground truth point of collection and the other is the predicted point. The *max - distance* indicates the upper bound of error, while the *min - distance* gives the error in the best case scenario. *max - distance* records an error of $\sqrt{2} \times scale_i^2$ even if both ground truth cell and predicted cell are the same and calculates the distance between the two furthest corners of the given cells. Conversely, *min - distance* gives an error of zero if the predicted cell and the ground truth cell are the same or are adjacent to each other, calculating the minimum distance between the two cells. For GEOLocate and the generative LLMs, we use the mean Simple Accuracy Error (SAE) between coordinate pairs. We also compare the methods on the percentage of predictions that lie within a 1km, 3km, 10km and $scale_i$ radius of the actual location.

■ **Table 1** Average distance errors and percentage of predictions within range of ground truth across the dataset.

Method		Average distance (km)	% acc@ 1km	% acc@ 3km	% acc@ 10km	% acc@ $scale_i$
GEOLocate _{text}		107.23	16.0	28.0	52.0	8.0
GEOLocate _{text+region}		107.23	16.0	28.0	52.0	8.0
ChatGPT _{text}		10.91	8.0	16.0	64.0	4.0
ChatGPT _{text+region}		10.12	8.0	16.0	68.0	
GPT-4o _{text}		155.82	4.0	16.0	40.0	8.0
GPT-4o _{text+region}		39.98	0	12.0	56.0	0
Our method	min	0.42	84.0	96.0	100	88.0
	max	2.16	24.0	80	100	0
	centroid	1.03	60.0	96.0	100	32.0

4.4 Results

Table 1 reports the performance of all methods tested. Both methods utilizing **GEOLocate** produced identical results, signaling that the region and country attributes do not contribute meaningfully to the georeferencing process. This may vary in other regions, such as the United States, where the state-based administrative system is more relevant as indicated in the documentation of GEOLocate. Out of the baselines, **ChatGPT_{text+region}** shows the best results with an average error of 10.12km. **ChatGPT_{text}** follows closely behind with no significant reduction in average distance. This indicates the LLM’s ability to disambiguate places to a high degree of accuracy even without the region or country information. **GPT-4o_{text}** produces the highest distance error. However, enhancing the prompt with the region, as in **GPT-4o_{text+region}**, significantly improves results. This suggests the LLM’s use of region for disambiguation. The stark difference in performance between the browser versions (ChatGPT_{text+region}, ChatGPT_{text}) and the same model accessed via the API (GPT-4o_{text}, GPT-4o_{text+region}) raise an important issue: the inability to browse the web in the API versions significantly hinders the quality of georeferencing. This is also observed in some of the reasoning provided by the model when producing the results. Versions with internet access are able to produce accurate coordinates for named places in the locality descriptions. This also leaves room for further improvement of the LLM based approaches. Providing precise and accurate locations for the named places may result in better quality. However, these improvements are not within scope of this paper.

Another interesting observation is the change of % acc at various distances. Although ChatGPT_{text+region} and ChatGPT_{text} produced lower errors (out of the baselines), the % acc@1km, and % acc@3km are worse than those of **GEOLocate** methods. Although able to correctly disambiguate the places and predict within the vicinity, all the LLM based approaches struggle to make a fine-grained prediction. This is to be expected as these methods only predict using point coordinates. Especially for large features such as rivers, mountains, and reserves, a point alone is an inadequate representation for an accurate georeferencing. Furthermore, these results indicate the LLM’s inability to take adequate consideration of the rich spatial relations commonly found in these locality descriptions.

The LMM we used for this experiment to test our approach is the OpenAI gpt-4o-2024-08-06 model accessed through their API. As previously discussed in Section 3.4.1, our prompt for the LMM does not limit the prediction of multiple grid cells. In our experiments, when the model predicts multiple cells, we only consider the first cell predicted. Our proposed

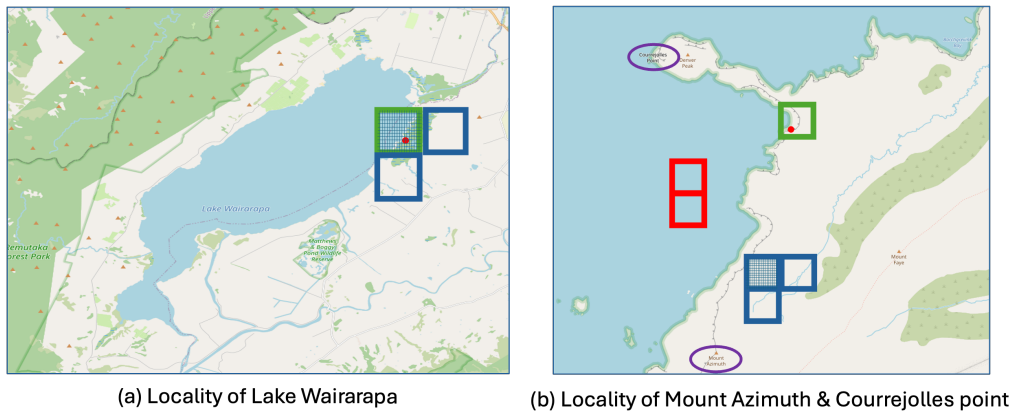


Figure 3 Map excerpts, their labels and their predictions for two locality descriptions: (a) J.K. Donald Wildlife Reserve, NE shore of L. Wairarapa – about 400m from lake & (b) Mount Azimuth, cliffs between Azimuth and Courrejolles Point near low point in ridge. The grid sizes for (a) and (b) are 1.88km and 0.7 km respectively. The red point indicates the exact point of collection. The green cell indicates the grid cell containing this point. The blue meshed cell indicates the first and primary cell predicted by the model and the other blue cells indicate the secondary predictions. The two place names mentioned in (b) are highlighted for visual clarity and the red cells indicate some of the cells considered during the reasoning of the model.

approach significantly outperforms the baselines. The centroid-distance of the LMM is an order of magnitude more accurate than the best-performing baseline. Max-distance, which is the upper bound for error given two grid cells, is also markedly lower than all baselines. This indicates our method’s ability to consider intricate spatial relations when producing georeferences. When considering a centroid-centroid distance, 60% of the predictions lie within 1km range of the actual location of collection. This level of accuracy is crucial when manually retrieving biological specimens. 32% of the predictions made by our multi-modal approach fall exactly in the correct grid cell as the original location. These results clearly demonstrate the significantly superior performance and usefulness of our grid-based multi-modal approach.

4.5 Discussion

4.5.1 Spatial extent and terrain understanding

A unique advantage of a multi-modal approach to georeferencing is its potential to understand spatial extents without being limited to simple coordinates. We analyzed the results to identify if the model is indeed capable of understanding extents of features. Figure 3(a) demonstrates an example where the model accurately identified the correct grid cell containing the point of collection. This is the map excerpt and prediction for the locality description shown in Figure 1. OSM did not find a match for J.K. Donald Wildlife Reserve and the model was restricted to only looking at the lake and its locality. Despite this, the model’s ability to correctly predict the grid cell demonstrates the model’s ability not only to identify the boundaries of the lake but also the distance from the border where the collection may have taken place (i.e. the “shore” in the locality description). Also of interest is the reasoning it produced for the prediction. The LLM response stated that it considers the green area that looks like a “vegetation patch” to be the J.K. Donald wildlife reserve. This shows the model’s ability to identify and reason with topographic features on the base map. Although

the LLM’s mentioned feature identity is questionable (as OSM’s name for that patch is Wairarapa Moana Wetland), this highlights a capability that could be highly beneficial for map-based spatial reasoning.

Figure 3(b) provides another similar example. In this case, the prediction is far from the actual collection location. However this is understandable when we analyse the locality description: *Mount Azimuth, cliffs between Azimuth and Courrejolles Point near low point in ridge*. Without contour lines or other altitude information, the phrase “low point in ridge” is indiscernible. What is of interest is the calculation the model made for “between”. The initial reasoning calculations made by the model predicted the cells marked in red as the cells that represent “between Azimuth and Courrejolles Point”. However, it later disregarded these cells in favour of the grids marked in blue. Although not explicitly stated, it seems to have avoided predicting a place in the ocean. This may also have been helped by the mention of an unnamed cliff. This ability of understanding terrain as shown in both examples opens the door to incorporating species-related habitat information into our approach. This could include characteristics such as whether a species inhabits land or water and even probabilistic heat maps on a species’ preferential ecosystem.

4.5.2 Linear Features

As mentioned earlier during the creation of the gridded map dataset, manual intervention was needed in the case of linear features. 9 out of the 25 samples contained linear features. Figure 4 demonstrates this issue, presenting two map excerpts for the following locality description: *“North Canterbury, Napenape Scenic Reserve, 3km south of mouth of Blythe River on coast.”*. Including the complete linear feature resulted in a vastly coarser grained map where the subsequently applied grid cells were 1.25km in scale. The map excerpt relevant for the accurate georeferencing would produce much finer grained cells of size 450m, allowing the model to not only pay attention to the river and the reserve but also differentiate grid cells based on whether they lie close to the coast or not. The proposed framework will benefit from further experiments on limiting the extent of the map especially with regard to linear features. A potential avenue is the exploration of distances to the other mentioned features and using these relations to limit the scope of the map.

Another observation on linear features was the vision encoder’s difficulty in comprehending the continuity of the linear features. Some confusion was observed when one road meets another at a junction but continues to be the same road after it. However, this can be remedied by custom labels placed at regular intervals of the linear feature.

4.5.3 Enhancing vision models’ map comprehension

Along with the confusion with linear features, we also noticed a tendency of the model to misrecognize the location of a feature using the label on the map instead of the icon or marker. This is contrary to findings in coarser grained maps [63]. These issues persist due to models like GPT-4o(Vision) not being specifically trained for map comprehension. Despite these inaccuracies, the performance of this zero-shot multi-modal approach is vastly superior to text only approaches. However, there is still space for improvement through fine-tuning, which would take into account the considerable variation in the forms of locality descriptions. The large numbers of natural history records collected from many different countries around the globe with detailed locality descriptions present an invaluable source of information to fine-tune (or perhaps even use during pre-training) vision models on map comprehension. Maps created using our framework can easily be annotated using

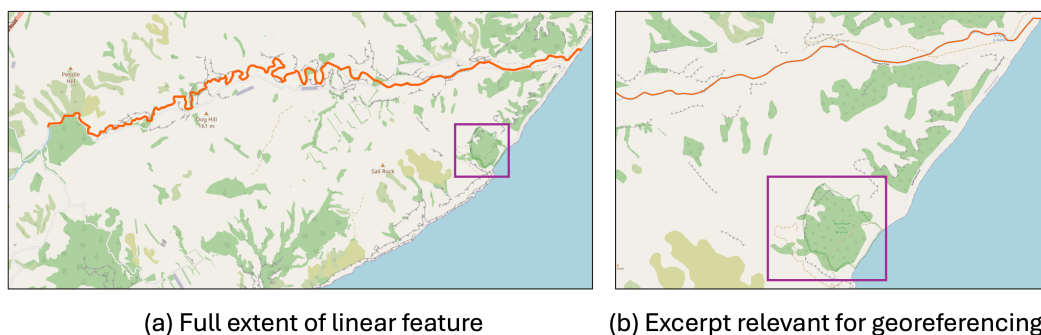


Figure 4 Two map excerpts for the same locality description. The inclusion of the full extent of the river (highlighted in red), as shown in (a) produces a much coarser map compared to (b). The Napenape Scenic Reserve is segmented in purple for visual clarity.

existing vision models: thus the framework could be used to create a version of the map with the point of collection prominently marked. Existing multi-modal models can then be used for the labelling (“Which grid cell contains the <Red Marker>?”) of these maps. These labels can subsequently be used for fine-tuning vision capabilities of other LMMs using the version of the map where the point of collection is removed. Alternatively, this can be used to pre-train open source vision encoders jointly with smaller open weight LLMs¹¹ to build LMMs specialized in map reading. This framework, of distantly supervised learning with cheap machine annotated data, can be regarded as analogous to masked language modeling or next sequence prediction for uni-modal language models.

5 Conclusion

This paper presents a novel method for georeferencing textual locality descriptions using LMMs to combine text understanding with map reading. The accuracy of this method is tested against existing tools and the current state-of-the-art LLMs where our method demonstrates greatly superior results. The distance error improves by an order of magnitude compared to the best baseline. Motivated by these results, a framework and workflow were designed to practically integrate LMMs for the task of georeferencing locality descriptions. Along with the model’s unique abilities and current shortcomings, the study also revealed avenues for future research that can be used to build powerful models capable of true map comprehension, taking one more step towards GeoAI.

References

- 1 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint*, 2023. [arXiv:2303.08774](#).
- 2 Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

¹¹Where the weights (parameters) of the LLM model are accessible

- 3 Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. doi:10.1109/ICCV.2015.279.
- 4 Arturo H Ariño. Approaches to estimating the universe of natural history collections data. *Biodiversity informatics*, 7(2), 2010.
- 5 Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. doi:10.48550/arXiv.2308.12966.
- 6 Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc., 2009.
- 7 Arthur D Chapman and John R Wiecezorek. Georeferencing best practices. Version 1.0, 2020. doi:10.15468/doc-gg7h-s853.
- 8 Hao Chen, Stephan Winter, and Maria Vasardani. Georeferencing places from collective human descriptions using place graphs. *Journal of Spatial Information Science*, 17:31–62, 2018. doi:10.5311/JOSIS.2018.17.417.
- 9 Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint*, 2025. doi:10.48550/arXiv.2501.17811.
- 10 Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023. URL: <https://jmlr.org/papers/v24/22-1144.html>.
- 11 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. arXiv:1810.04805.
- 12 Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint*, 2023. arXiv:2303.03378.
- 13 Jianzhou Feng, Ganlin Xu, Qin Wang, Yuzhuo Yang, and Lei Huang. Note the hierarchy: Taxonomy-guided prototype for few-shot named entity recognition. *Information Processing & Management*, 61(1):103557, 2024. doi:10.1016/J.IPM.2023.103557.
- 14 Jenny Rose Finkel, Trond Grenager, and Christopher D Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL'05)*, pages 363–370, 2005. doi:10.3115/1219840.1219885.
- 15 Milan Gritta, Mohammad Taher Pilehvar, and Nigel Collier. Which Melbourne? augmenting geocoding with maps. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1285–1296, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi:10.18653/v1/P18-1119.
- 16 Milan Gritta, Mohammad Taher Pilehvar, Nut Limsopatham, and Nigel Collier. What's missing in geographical parsing? *Language Resources and Evaluation*, 52:603–623, 2018. doi:10.1007/S10579-017-9385-8.
- 17 Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint*, 2025. arXiv:2501.12948.
- 18 Qinghua Guo, Yu Liu, and John Wiecezorek. Georeferencing locality descriptions and computing associated uncertainty using a probabilistic approach. *International Journal of Geographical Information Science*, 22(10):1067–1090, 2008. doi:10.1080/13658810701851420.

- 19 Andreas Hackeloeer, Klaas Klasing, Jukka M Krisp, and Liqiu Meng. Georeferencing: a review of methods and applications. *Annals of GIS*, 20(1):61–69, 2014. doi:10.1080/19475683.2013.868826.
- 20 Linda L Hill. *Georeferencing: The geographic associations of information*. Mit Press, 2009.
- 21 Hartwig H Hochmair, Levente Juhász, and Takoda Kemp. Correctness comparison of chatgpt-4, gemini, claude-3, and copilot for spatial tasks. *Transactions in GIS*, 28(7):2219–2231, 2024. doi:10.1111/TGIS.13233.
- 22 Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, et al. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, page ocad259, 2024.
- 23 A. Khan, M. Vasardani, and S. Winter. Extracting spatial information from place descriptions. In *COMP '13 ACM SIGSPATIAL International Workshop on Computational Models of Place*, pages 62–69, New York, NY, USA, 2013. Association for Computing Machinery. doi:10.1145/2534848.2534857.
- 24 Yoonsik Kim, Moonbin Yim, and Ka Yeon Song. Tablevqa-bench: A visual question answering benchmark on multiple table domains. *arXiv preprint arXiv:2404.19205*, 2024. doi:10.48550/arXiv.2404.19205.
- 25 Parisa Kordjamshidi, Martijn Van Otterlo, and Marie-Francine Moens. Spatial role labeling: Towards extraction of spatial relations from natural language. *ACM Transactions on Speech and Language Processing (TSLP)*, 8(3):1–36, 2011. doi:10.1145/2050104.2050105.
- 26 John R Lawson, Joseph E Trujillo-Falcón, David M Schultz, Montgomery L Flora, Kevin H Goebbert, Seth N Lyman, Corey K Potvin, and Adam J Stepanek. Pixels and predictions: Potential of gpt-4v in meteorological imagery analysis and forecast communication. *Artificial Intelligence for the Earth Systems*, 4(1):240029, 2025.
- 27 Jochen L Leidner, Gail Sinclair, and Bonnie Webber. Grounding spatial named entities for information extraction and question answering. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references*, pages 31–38, 2003.
- 28 Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. A survey on deep learning for named entity recognition. *IEEE transactions on knowledge and data engineering*, 34(1):50–70, 2020.
- 29 Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. URL: <https://proceedings.mlr.press/v162/li22n.html>.
- 30 Menglin Li, Kwan Hui Lim, Teng Guo, and Junhua Liu. A transformer-based framework for poi-level social post geolocation. In Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo, editors, *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part I*, volume 13980 of *Lecture Notes in Computer Science*, pages 588–604. Springer, 2023. doi:10.1007/978-3-031-28244-7_37.
- 31 Krzysztof Lipka, Dariusz Gotlib, and Kamil Choromański. The use of language models to support the development of cartographic descriptions of a building’s interior. *Applied Sciences*, 14(20):9343, 2024.
- 32 Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint*, 2024. arXiv:2412.19437.
- 33 Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.
- 34 Y. Liu, Q. H. Guo, J. Wiecezorek, and M. F. Goodchild. Positioning localities based on spatial assertions. *International Journal of Geographical Information Science*, 23(11):1471–1501, 2009. doi:10.1080/13658810802247114.

- 35 Oswaldo Ludwig, Xiao Liu, Parisa Kordjamshidi, and Marie-Francine Moens. Deep embedding for spatial role labeling. *arXiv preprint arXiv:1603.08474*, 2016. [arXiv:1603.08474](#).
- 36 A Marcer, Quentin Groom, Elspeth Haston, and Francesc Uribe. Natural history collections georeferencing survey report. *Current georeferencing practices across institutions worldwide*. Zenodo, 2021.
- 37 Fernando Melo and Bruno Martins. Automated geocoding of textual documents: A survey of current approaches. *Transactions in GIS*, 21(1):3–38, 2017. [doi:10.1111/TGIS.12212](#).
- 38 David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. In *Named Entities: Recognition, classification and use*, pages 3–28. John Benjamins publishing company, 2009.
- 39 Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. Improving multimodal datasets with image captioning. *Advances in Neural Information Processing Systems*, 36:22047–22069, 2023.
- 40 GPT OpenAI. 4v (ision) system card. *preprint*, 2023.
- 41 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- 42 Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- 43 Nelson E Rios and Henry L Bart Jr. Geolocate - software for georeferencing natural history data, Year of Access or Publication. Accessed: 10 Feb. 2025. URL: <https://www.geo-locate.org>.
- 44 Yves Scherrer, Nikola Ljubešić, et al. Social media variety geolocation with geobert. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 135–140. The Association for Computational Linguistics, 2021.
- 45 Florian Schneider and Sunayana Sitaram. M5—a diverse benchmark to assess the performance of large multimodal models across multilingual and multicultural vision-language tasks. *arXiv preprint arXiv:2407.03791*, 2024. [doi:10.48550/arXiv.2407.03791](#).
- 46 Pavel Serdyukov, Vanessa Murdock, and Roelof Van Zwol. Placing flickr photos on a map. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 484–491. ACM, 2009. [doi:10.1145/1571941.1572025](#).
- 47 Lihardo Faisal Simanjuntak, Rahmad Mahendra, and Evi Yulianti. We know you are living in bali: Location prediction of twitter users using bert language model. *Big Data and Cognitive Computing*, 6(3):77, 2022. [doi:10.3390/BDCC6030077](#).
- 48 Xinshuai Song, Weixing Chen, Yang Liu, Weikai Chen, Guanbin Li, and Liang Lin. Towards long-horizon vision-language navigation: Platform, benchmark and method. *arXiv preprint arXiv:2412.09082*, 2024. [doi:10.48550/arXiv.2412.09082](#).
- 49 Kristin Stock, Kalana Wijegunaratna, Christopher B Jones, Hone Morris, Pragyan Das, David Medyckyj-Scott, and Brandon Whitehead. The biowhere project: unlocking the potential of biological collections data. *GI_Forum*, 11(1):3–21, 2023.
- 50 Jianlin Su, Ahmed Murtadha, Shengfeng Pan, Jing Hou, Jun Sun, Wanwei Huang, Bo Wen, and Yunfeng Liu. Global pointer: Novel efficient span-based approach for named entity recognition. *arXiv preprint arXiv:2208.03054*, 2022. [doi:10.48550/arXiv.2208.03054](#).
- 51 Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. Anaphora and coreference resolution: A review. *Information Fusion*, 59:139–162, 2020. [doi:10.1016/J.INFFUS.2020.01.010](#).
- 52 Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint*, 2023. [arXiv:2307.09288](#).
- 53 Marieke van Erp, Robert Hensel, Davide Ceolin, and Marian Van der Meij. Georeferencing animal specimen datasets. *Transactions in GIS*, 19(4):563–581, 2015. [doi:10.1111/TGIS.12110](#).

- 54 Marieke van Erp, Robert Hensel, Davide Ceolin, and Marian Van der Meij. Georeferencing animal specimen datasets. *Transactions in GIS*, 19(4):563–581, 2015. doi:10.1111/TGIS.12110.
- 55 Olivier Van Laere, Steven Schockaert, Vlad Tanasescu, Bart Dhoedt, and Christopher B. Jones. Georeferencing wikipedia documents using data from social media sources. *ACM Trans. Inf. Syst.*, 32(3), July 2014. doi:10.1145/2629685.
- 56 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.
- 57 Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. Gpt-re: In-context learning for relation extraction using large language models. *arXiv preprint arXiv:2305.02105*, 2023. doi:10.48550/arXiv.2305.02105.
- 58 Jimin Wang and Yingjie Hu. Enhancing spatial and textual analysis with eupeg: An extensible and unified platform for evaluating geoparsers. *Transactions in GIS*, 23(6):1393–1419, 2019. doi:10.1111/tgis.12579.
- 59 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- 60 Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*, 2023. doi:10.48550/arXiv.2302.11382.
- 61 John Wiecek, Qinghua Guo, and Robert Hijmans. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International journal of geographical information science*, 18(8):745–767, 2004. doi:10.1080/13658810412331280211.
- 62 Benjamin Wing and Jason Baldridge. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 955–964, 2011. URL: <https://aclanthology.org/P11-1096/>.
- 63 Jinwen Xu and Ran Tao. Map reading and analysis with gpt-4v (ision). *ISPRS International Journal of Geo-Information*, 13(4):127, 2024. doi:10.3390/IJGI13040127.
- 64 Chris Yesson, Peter W Brewer, Tim Sutton, Neil Caithness, Jaspreet S Pahwa, Mikhaila Burgess, W Alec Gray, Richard J White, Andrew C Jones, Frank A Bisby, et al. How global is the global biodiversity information facility? *PloS one*, 2(11):e1124, 2007.
- 65 Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint*, 2024. arXiv:2404.16006.
- 66 Meishan Zhang, Bin Wang, Hao Fei, and Min Zhang. In-context learning for few-shot nested named entity recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10026–10030. IEEE, 2024. doi:10.1109/ICASSP48485.2024.10446653.
- 67 Qianqian Zhang, Mengdong Chen, and Lianzhong Liu. A review on entity relation extraction. In *2017 second international conference on mechanical, control and computer engineering (ICMCCE)*, pages 178–183. IEEE, 2017.
- 68 Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022. doi:10.48550/arXiv.2210.03493.
- 69 Xufeng Zhao, Mengdi Li, Wenhao Lu, Cornelius Weber, Jae Hee Lee, Kun Chu, and Stefan Wermter. Enhancing zero-shot chain-of-thought reasoning in large language models through logic. *arXiv preprint arXiv:2309.13339*, 2023. doi:10.48550/arXiv.2309.13339.

- 70 Zexuan Zhong and Danqi Chen. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, 2021. doi:10.18653/V1/2021.NAACL-MAIN.5.
- 71 Mengjie Zhou, Liu Liu, Yiran Zhong, and Andrew Calway. Geolocation on cartographic maps with multi-modal fusion. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5589–5596. IEEE, 2024. doi:10.1109/IROS58592.2024.10801404.

Assessing Map Reproducibility with Visual Question-Answering: An Empirical Evaluation

Eftychia Koukouraki¹  

Institute for Geoinformatics, University of Münster, Germany

Auriol Degbelo  

Chair of Geoinformatics, TU Dresden, Germany

Christian Kray  

Institute for Geoinformatics, University of Münster, Germany

Abstract

Reproducibility is a key principle of the modern scientific method. Maps, as an important means of communicating scientific results in GIScience and across disciplines, should be reproducible. Currently, map reproducibility assessment is done manually, which makes the assessment process tedious and time-consuming, ultimately limiting its efficiency. Hence, this work explores the extent to which Visual Question-Answering (VQA) can be used to automate some tasks relevant to map reproducibility assessment. We selected five state-of-the-art vision language models (VLMs) and followed a three-step approach to evaluate their ability to discriminate between maps and other images, interpret map content, and compare two map images using VQA. Our results show that current VLMs already possess map-reading capabilities and demonstrate understanding of spatial concepts, such as cardinal directions, geographic scope, and legend interpretation. Our paper demonstrates the potential of using VQA to support reproducibility assessment and highlights the outstanding issues that need to be addressed to achieve accurate, trustworthy map descriptions, thereby reducing the time and effort required by human evaluators.

2012 ACM Subject Classification Information systems → Question answering; Computing methodologies → Spatial and physical reasoning; Human-centered computing → Geographic visualization; Applied computing → Cartography

Keywords and phrases map comparison, computational reproducibility, visual question answering, large language models, GeoAI

Digital Object Identifier 10.4230/LIPIcs.GIScience.2025.13

Supplementary Material *Software (Source Code)*: <https://doi.org/10.17605/OSF.IO/W4BQG>

Dataset: <https://doi.org/10.17605/OSF.IO/W4BQG>

Other (Intermediate Results): <https://doi.org/10.17605/OSF.IO/W4BQG>

Funding *Eftychia Koukouraki*: Erasmus+ programme, Erasmus Mundus action 2021-2027, project number 101049796.

Auriol Degbelo: German Research Foundation, NFDI4Earth, project number 460036893.

1 Introduction and Background

Maps play a key role in information visualisation, serving as an essential tool for communicating insights from geographic and spatial data. Geographic maps are published in various outlets, from scientific journals to newspapers, which makes them accessible to a wide range of audiences. Maps in scientific outlets, in particular, should represent the world truthfully and accurately within known limits of precision [14], and ideally be reproducible in order to provide reliable evidence for findings and facilitate the communication of science

¹ Corresponding author



© Eftychia Koukouraki, Auriol Degbelo, and Christian Kray;
licensed under Creative Commons License CC-BY 4.0

13th International Conference on Geographic Information Science (GIScience 2025).

Editors: Katarzyna Sila-Nowicka, Antoni Moore, David O'Sullivan, Benjamin Adams, and Mark Gahegan;
Article No. 13; pp. 13:1–13:12



Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

to society. Take, for instance, the field of environmental sciences, where climate change debates have grown increasingly polarised. Maps can be powerful tools in these discussions, but they can also be used to promote competing agendas. If inaccurate or misleading, maps can lead to serious consequences, including poor decision making and hindered climate action. This example illustrates a larger point: the need for transparent and reproducible map making standards that can be applied across domains to support informed decision making and maintain scientific integrity. Current practices of overpublishing that favour quantity over quality in research publications [1], combined with the explosion of generative artificial intelligence (AI), have made reproducibility increasingly important for establishing the credibility of published research, for verifying results, and for enabling current studies to be reused and built upon.

Reproducibility is defined as the ability to reach the same results previously obtained by other researchers after repeating a scientific experiment based on the same data and methods [18]. This can only be achieved if the data and software that underpinned a study are transparent and accessible, but even then it is often not possible in practice to achieve exactly the same results as the original study, especially when it comes to reproducing visualisations. Inadequate documentation, the use of different software packages, and the reliance on implicit system configurations are common causes of discrepancies between the reproduced results and the original findings [15, 16, 19]. To ensure that a study is reproducible, the reproduced results must be evaluated against the original results [12, 15, 17, 24].

Basing the evaluation of reproduced results on numerical values is generally a straightforward process: if all numbers are identical, the reproduction is considered successful. However, visualised results, e.g. in the form of diagrams or maps, are easier to grasp for human observers, but pose several challenges when used to assess reproducibility. Variations in graph curves, missing key numbers, and different aspect ratios can make it difficult for readers to determine if reproduced figures accurately reflect the original results, even when the numerical data is identical [15]. In addition, an increased effort required for map reading can negatively impact the evaluator's assessment of the success of the map reproduction [17]. Therefore, computational support is essential for assessing reproduced maps in order to increase efficiency and accuracy, as well as to facilitate the examination of geovisualisations illustrating complex datasets. Besides, multiple maps can be derived from a single dataset during scientific exploration. Nonetheless, reproducible map making focuses on creating a faithful visual copy of an original map, without introducing any significant variations that alter the map's interpretation [16]. Hence, only two maps are of interest during reproducibility assessment: the original and its copy. The goal is to identify similarities or differences between them - using supporting data, software, and documentation - without concern for the map's ontological status (e.g. as truth, social construct, or mappings [13]). Of course, graphical excellence and graphical integrity [21] remain essential to ensure that both the original and the copy do not distort effects in the underlying source data.

Question-answering (QA) can serve as a method for extracting and evaluating map content [20]. Visual question-answering (VQA), the computer vision task of teaching machines to comprehend the content of a picture and to answer questions about it in natural language, can now be supported by vision language models (VLMs), which are multimodal large language models (LLMs) capable of processing and understanding both text and image. For example, Bendeck and Stasko [3] explored the potential of VLMs for the visual interpretation of charts, confirming their capabilities while also highlighting their current limitations in this task. Thus, we can infer that maps, as a specialised type of chart with explicit spatial relationships between the depicted elements, could also benefit from these advancements.

In order to investigate to what degree this is true in practice, we examined the ability of five state-of-the-art models to support tasks related to map reproduction assessment. Our goal was to assess whether VLMs can assist an independent reproducing researcher or a reproducibility reviewer to verify that a map has been successfully reproduced. In particular, we examined three tasks: map discrimination (distinguishing between map and non-map images), map interpretation (answering questions about a map image accurately), and map comparison (assessing the similarity of two maps based on a set of questions).

Recent studies in GIScience have explored the understanding of LLMs for spatial concepts. For instance, Ji and Gao [11] evaluated the ability of LLMs (GPT-2, BERT) to represent geometries and their spatial relations using LLM-generated embeddings. The results showed the potential of LLMs to capture geometry types and spatial relations, while there is room for improvement in estimating numerical values and retrieving spatially related objects. The capacity of LLMs for spatial reasoning was also confirmed by Cohn and Blackwell [6]. However, they concluded that LLMs are not reliable for drawing conclusions about cardinal directions and perform better in factual recall tasks rather than in spatial reasoning tasks. Hojati and Feick [9] tested the performance of various LLMs in answering spatial questions and providing methodological steps for arriving at each answer, both in natural language and in SQL. Feng et al. [7] connected the prompt to an external knowledge base to develop a Geographic Question Answering (GeoQA) pipeline, thereby extending the capabilities of LLMs. Moving from text-only to multimodal input (i.e. text accompanied by images), Xu and Tao [23] found that GPT-4V could retrieve information and perform basic analysis tasks with maps. Griffin and Robinson [8] used the ChatGPT prompt to generate accessibility descriptions for map input. While the aforementioned studies demonstrate encouraging results, multimodal input has yet to be systematically tested for spatial concepts.

Our research addresses this challenge, setting the context of QA in relation to the assessment of map reproducibility. The key contribution of this paper is the empirical evaluation of five state-of-the-art VLMs for three key tasks related to map reproducibility assessment: map discrimination, interpretation and comparison. Our findings demonstrate that, albeit with certain limitations, VLM-enabled VQA can streamline the verification of reproduced scientific results displayed on maps. In addition to its benefits in automating reproducibility assessment, map VQA also has the potential to improve accessibility as it opens up new possibilities for visually impaired readers to access information in figures.

2 Experimental Design

To examine the interpretation capabilities of VLMs for maps, we selected five state-of-the-art VLMs based on performance and diversity. Specifically, we considered the models with the highest scores on the vision leaderboard in the Chatbot Arena [5], ensuring that no two models were from the same provider (e.g. Google or OpenAI). We did not consider models that might be subject to rate limits or withdrawn without prior notice, such as those labelled as experimental or preview. The selection was done at the beginning of January 2025 and this led to the following five models: Gemini 2.0 Flash-001, GPT-4o (2024-11-20), Claude 3.5 Sonnet (20240620), Pixtral Large (latest), and Qwen-VL-Max. We narrowed down the scope of this study by focusing solely on thematic maps and followed a three-step approach to evaluate the map reasoning skills of the selected VLMs:

Step 1 – Map Discrimination The ability to distinguish between different types of charts - between maps and non-maps in this case - is necessary for automating the reproducibility assessment of visualisations. We considered this step a prerequisite for confirming that

the VLM understands the concept of a geographic map and can therefore be used to automate subsequent tasks related to the reproducibility assessment of geovisualisations. Hence, we tested the ability of the models to differentiate between maps and other types of charts. We assembled a dataset of 40 images, consisting of 20 maps and 20 charts of other types, including pie, line, bar and point charts, and posed the question *Is this image showing a map?*. To account for the diversity of maps encountered in different outlets, we sourced maps from Our World in Data, which targets broader audiences, and from the scientific Journal of Maps. The selected maps cover a range of geographic scopes, from regional to global; different layout and legend styles; and different applications, from geological to socio-economic indicators. All the images we sourced were licensed under CC BY.

Step 2 – Map Interpretation Extracting and evaluating information from maps is essential for assessing the equivalence between one map and another. Therefore, we tested the VLM’s ability to read and interpret geographic maps. We asked eight questions about map interpretation on the map subset from the map discrimination task, each question addressing one of the following dimensions: map type, spatial scale, geographic scope, orientation, visualised data, symbology, legend recognition, and legend-data consistency.

Step 3 – Map Comparison The final step in assessing reproducibility is to compare the reproduced result with the original, as mentioned in Section 1. Therefore, we evaluated the VLMs’ map comparison capabilities. For this step, we used a dataset of 20 maps that differ from each other in only one dimension, such as orientation, symbology, or legend, to assess whether the models can identify subtle visual nuances that are relevant in geographic information visualisation. We provided two maps as input to the models and asked six questions about their differences, following the guidelines on the importance of visual differences in assessing map reproductions provided by [17]. The questions addressed similarities in the topic, geographic extent, orientation, positions of the visualised data, legend, and symbology. All questions were formulated to be answerable with *yes* or *no*, so that a human evaluator could quickly skim through the automated responses and determine whether any significant differences were identified.

The aforementioned steps were implemented using the models’ APIs in Python scripts. In the API calls, we set the models’ attributes *temperature* to “0” and, if applicable, *seed* to the same random integer (“123”) to make the model as deterministic as possible. We also set the maximum number of tokens in the model’s response to 128, assuming that this number of tokens should be sufficient to provide a focused answer. If the model exhausted this limit for most answers, we reran the test and set a new maximum number of 160 tokens. We did not extend the token limit beyond this number. Additionally, we measured the time each model took to respond to each prompt and calculated the mean completion time per output token.

A sample of the dataset for all three steps is shown on Figure 1. The entire dataset and the scripts created for this experiment can be accessed at <https://doi.org/10.17605/OSF.IO/W4BQG>.

Evaluation

We evaluated the accuracy of the map discrimination task based on the model’s ability to correctly answer *yes* or *no*, without further analysing the responses. For the map interpretation and comparison tasks, we evaluated the models’ constrained response accuracy by classifying an answer as correct if all the information provided within the specified token limit was accurate; otherwise, it was classified as incorrect. This metric indicates the model’s

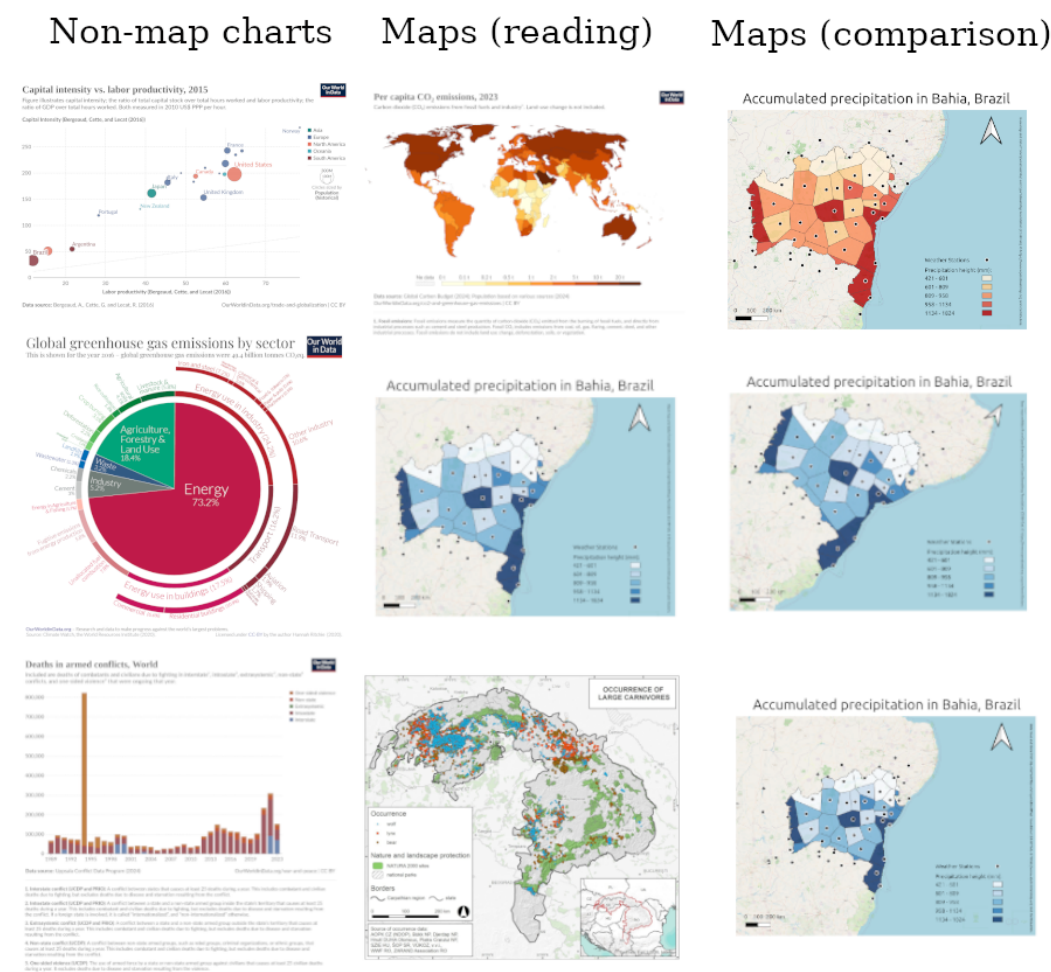


Figure 1 A selection of images from the dataset we compiled. Images are sourced from Our World in Data (<https://ourworldindata.org/data>) and Vlkova et al. [22] under the CC BY license, and our own creations. The map comparison figures were generated by systematically varying one dimension (e.g. color, orientation, or symbology) of an existing map image.

ability to provide an accurate answer to the question within the specified token limit, without including any false information. For example, if the model gave a correct answer but provided a false justification or included any incorrect information along with the answer, we marked it as incorrect. Similarly, if the model provided several true facts but failed to explicitly answer the question within the token limit, we also marked it as incorrect. Our guiding principle for the evaluation was whether the model could be trusted to provide accurate information without requiring our supervision. During the evaluation process, we kept a log of issues that arose and could help further characterise the use of VLMs for this purpose, but that could not be quantified in terms of correct/incorrect percentages. We also did this to gain a qualitative impression of the models' strengths and weaknesses.

Prototype

As mentioned in Section 1, automated tests are a desirable asset for map reproducibility assessment. With the best performing model, we built a browser-based tool that allows users to upload two map images, run the evaluation process, and determine if the second

image has been successfully reproduced. The evaluation process is based on the questionnaire we created for the map comparison task. Additionally, we implemented a simple overall evaluation function that counts the number of *yes* and *no* answers, and returns a successful status if more than half of the questions were answered with *yes* and unsuccessful otherwise. We also integrated the calculation of cosine similarity based on the image embeddings of the two input images as an initial quantitative indicator of their similarity. The code of the prototype can be found at <https://doi.org/10.17605/OSF.IO/W4BQG>.

3 Results

We ran all the experiments from the same Ethernet cable endpoint, which theoretically provides a 1000 Mbps Internet connection. In practice, we measured 936.71–937.38 Mbps for download and 874.27–933.74 Mbps for upload. We found the average completion time per output token, from shortest to longest, to be as follows: Pixtral Large (0.06 seconds), Gemini 2.0 Flash (0.08 seconds), Qwen-VL-Max (0.10 seconds), Claude Sonnet 3.5 (0.11 seconds), and GPT-4o (0.15 seconds). However, Gemini 2.0 Flash had the fastest overall completion time, as it provided shorter answers compared to the other models.

In the map discrimination task, all models were able to differentiate perfectly between maps and non-maps. It is worth mentioning that one of the geographic maps in this dataset included pie charts illustrating the ratio of mountain area to land surface for each continent, and all five models correctly classified this image as a map. It should also be noted that Qwen-VL-Max rejected nine of the 20 map images used as input for the map discrimination and map interpretation tasks. The error message returned was: “Input data may contain inappropriate content.” We were unable to identify any pattern related to map topic, geographic extent, or image resolution.

For the map interpretation task, we found the constrained response accuracy, from highest to lowest, to be as follows: Gemini 2.0 Flash (80%), GPT-4o (77%), Claude Sonnet 3.5 (76%), Qwen-VL-Max (69%), and Pixtral Large (58%). The lower performance of Pixtral Large is mostly due to an inability to give concise answers within the token limit, rather than providing factually inaccurate information. Qwen-VL-Max and Pixtral Large appear to rely heavily on Optical Character Recognition (OCR), as they seem to repeatedly use the text extracted from the image in their responses. This, combined with the text generation module, can lead to vague answers. Pixtral Large also tends to continue describing the entire image even after answering the question. The constrained response accuracy per question for the map interpretation task is shown in Table 1. We can observe that almost all models performed worst on the question *What is the spatial scale of the map?*. We accepted answers related to the scale bar as correct; however, the models often ignored the scale bar, misinterpreted it, or referred to the geographic extent instead. Conversely, the models achieved the highest average constrained response accuracy on the question regarding the geographic scope.

The models were able to identify and distinguish between several types of maps beyond thematic, including choropleth, topographic, tectonic, proximity, land cover and habitat suitability maps. GPT-4o provided the most diverse responses to this question. All models were able to identify inset maps, although they were not explicitly asked to do so. Moreover, the models are already performing some level of fact-checking, such as identifying the location of the highest mountain peaks. The generative nature of VLMs is also evident, as they tend to elaborate on aspects that were not the subject of the question. Gemini 2.0 Flash exhibited this behaviour the least.

To answer the question *What data are visualised on the map?*, the models essentially parsed and repeated the legend. They showed a good understanding of what a legend is and were able to recognise different legend formats. However, mapping visual symbols to their corresponding values is not always straightforward, especially in horizontal legends where each colour represents a range of values. We found Claude Sonnet 3.5 to be particularly effective at legend interpretation, providing many details.

■ **Table 1** Constrained response accuracy per question for the map interpretation task.

	Gemini 2.0 Flash	GPT-4o	Claude Sonnet 3.5	Pixtral Large	Qwen-VL- Max	Average per question
What type of map is this?	90%	80%	65%	55%	91%	76%
What is the geographic scope of the map?	95%	95%	70%	75%	82%	83%
What is the orientation of the map?	95%	90%	90%	85%	45%	81%
What data are visualised on the map?	90%	70%	95%	80%	73%	82%
What symbols are used to visualise the data on the map?	80%	65%	80%	45%	64%	67%
Does this map contain a legend?	70%	95%	85%	50%	82%	76%
Is the legend consistent with the visualised data?	90%	90%	75%	40%	64%	72%
What is the spatial scale of the map?	30%	30%	50%	35%	55%	40%
Average per model	80%	77%	76%	58%	69%	

For the map comparison task, we found the constrained response accuracy, from highest to lowest, to be as follows: GPT-4o (86%), Gemini 2.0 Flash (85%), Qwen-VL-Max (81%), Pixtral Large (74%), and Claude Sonnet 3.5 (73%). The constrained response accuracy per question for the map comparison task is shown in Table 2. We can observe that the performance in this task is better than in map interpretation. This could either be because the second image acts as additional context or reference, helping the model to provide accurate answers, or because the maps used for this task are less complex. The models achieved the highest average accuracy for the question on legend similarity, further reinforcing the impression that VLMs can effectively identify the map legend as a distinct object. All models scored lowest on the question *Do the two maps visualise the same data in the same positions?*. Our dataset included a map with data points shifted by several pixels compared to the original, but no model identified the difference. Claude Sonnet 3.5 responded that there was a difference in the data positions, but justified its answer by mentioning a difference in the distribution of colours. Moreover, when we presented two maps that show the same data but differ slightly in geographic extent (i.e. one map looks “zoomed in” compared to the other), Pixtral Large and GPT-4o interpreted this difference as a change in the visualised data pattern. This suggests they may be counting pixels rather than using object-based area quantification.

All models detected a difference in the units of measurement in the legend (cm instead of mm). GPT-4o, Pixtral Large, and Claude Sonnet 3.5 correctly identified a difference in the font, while GPT-4o and Gemini 2.0 Flash detected a change in the base map. All of these differences were detected by the models without explicitly asking for them in the prompt.

■ **Table 2** Constrained response accuracy per question for the map comparison task.

	Gemini 2.0 Flash	GPT-4o	Claude Sonnet 3.5	Pixtral Large	Qwen-VL- Max	Average per question
Are these maps about the same topic?	100%	100%	63%	79%	74%	83%
Do the maps have the same geographic extent?	74%	84%	68%	74%	89%	78%
Do the maps have the same orientation?	89%	79%	84%	74%	79%	81%
Do the two maps visualise the same data in the same positions?	74%	68%	58%	53%	63%	63%
Do the two maps have the same legend?	89%	95%	89%	74%	89%	87%
Do the two maps use the same symbols for the visualised data?	84%	89%	74%	89%	89%	85%
Average per model	85%	86%	73%	74%	81%	

4 Discussion and Outlook

Our results show that VQA is a promising tool for assessing map reproducibility. No model performed equally well on all questions, but the accuracy values obtained during the evaluation suggest capabilities to assess the content of a reproduced map that go beyond pixel-wise comparison. Another advantage of using VQA to assess reproducibility is its independence from specific tools, as it is only the data format of the final cartographic product that matters and not whether we have used scripts or desktop GIS to produce it. The use of VQA for content-based map comparison offers a new approach to assessing the equivalence of geovisualisations, not only in the context of reproducibility, but also in other scenarios, such as creating equivalent visualisations for different audiences (e.g. the scientific community, policy makers, the general public).

Questions where the response accuracy values are particularly low (Tables 1, 2) indicate areas for future research so that the models can come to the point where they can be confidently used in automated assessment workflows. Also, the reasons that affect the performance of a model (e.g. impact of the number of parameters, training process) should be systematically investigated in future work before its integration into these workflows.

Furthermore, two key conceptual issues must be addressed before integrating a VQA approach into automated systems. First, ensuring transparency throughout the entire assessment process is essential, which poses a challenge when working with VLMs/LLMs. If integrated into an automated assessment system, a model should be explainable to ensure fairness in automated decisions and to promote trust [2]. At the moment, the best model is Gemini 2.0 Flash, based on both speed and constrained response accuracy. However, relying on closed-source, proprietary models for such tasks contradicts the principles of open science. An automated reproducibility assessment system should itself be verifiable before it is used to verify scientific outcomes. To achieve this, we need open-source models with better performance. Another issue to resolve before automating the reproducibility evaluation process is determining the threshold for success. In this paper, we have based this evaluation on the similarity of the reproduced map to the original. While this comparison is necessary to confirm reproducibility, it is not sufficient on its own; factors such as the

accessibility of materials and the computational effort required for reproduction also indicate how reproducible a study is. While moving beyond pixel-based comparisons is a step forward, the question remains: how should reproducibility and reproduction success be quantified?

Future research should continue to advance our understanding of map comparison and VQA capabilities. One potential avenue for exploration involves investigating alternative comparison strategies other than the yes/no question approach that we followed. For example, we could perform text similarity computations on the answers, and examine additional comparison dimensions such as the units of measurement or the basemap. The set of questions used in this study was deliberately kept simple in order to have a consistent evaluation of the VLMs' responses across the different maps in our dataset and to establish an initial baseline for VLM evaluation. Developing more sophisticated and context-specific questions is part of our future work. Additionally, exploring the ability to accurately retrieve specific data values from different positions on the map presents another promising area for future research. It is also worth investigating the extent to which VLM responses are based on the textual elements on the maps and how well VQA would perform on maps with no or very little text.

Overall, the ability to get accurate descriptions of maps with VQA is remarkable not only for map reproducibility assessment, but also because it opens up new ways for visually impaired people to access information for the first time. It is also a step towards the democratisation of science, where VQA can be used by the public to get explanations of scientific geovisualisations [4].

Limitations

There are several limitations to our work. Firstly, we set the maximum number of output tokens to 128, with an option to extend it to 160. This means that the models might have hallucinated more (i.e. presented false information as fact [10]) or might have come to a different conclusion if we had allowed a higher limit. It is necessary to assess the sensitivity of the results to the maximum number of output tokens by setting different limits, evaluating the outcomes, and determining whether the results remain consistent across different limits. Furthermore, constrained response accuracy is only an initial measure of the models' performance. We did not develop specific metrics for conciseness, focus, or completeness, only qualitative notes were taken during the evaluation. Finally, the comparison task focused on maps varying along a single dimension, as mentioned in Section 2. The performance of VLMs on maps that differ across multiple dimensions, which adds complexity to this task, remains to be tested.

5 Summary

In this paper, we investigated the ability of five popular VLMs (Gemini 2.0 Flash, GPT-4o, Claude Sonnet 3.5, Pixtral Large, Qwen-VL-Max) to discriminate, interpret, and compare geographic maps using VQA. We compiled a set of 40 chart images (20 maps and 20 charts of other types) to test whether the VLMs can distinguish between maps and non-maps. Subsequently, we evaluated the VLMs using only the map images by asking questions covering eight dimensions of map interpretation. After confirming the potential of these models for interpreting geographic maps, we proceeded to evaluate their map comparison capabilities by providing two maps as input and asking questions about their identified differences across six dimensions relevant to assessing map reproduction [17]. For the comparison task, we used 20 maps that differ in only one dimension. While preliminary, our results show that

all five VLMs already possess spatial understanding and map reading skills. Our next steps in this line of research will aim to improve the models' performance and to develop more sophisticated strategies for comparing maps and quantifying their differences. Ultimately, we are working towards integrating VQA into systems that automate map reproduction assessment and support scientific fact-checking, enabling reproducibility reviewers to quickly verify scientific results.

References

- 1 Andrew R. Akbashev and Sergei V. Kalinin. Tackling overpublishing by moving to open-ended papers. *Nature Materials*, 22(3):270–271, March 2023. Publisher: Nature Publishing Group. doi:10.1038/s41563-023-01489-1.
- 2 Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bannetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, June 2020. doi:10.1016/j.inffus.2019.12.012.
- 3 Alexander Bendeck and John Stasko. An Empirical Evaluation of the GPT-4 Multimodal Language Model on Visualization Literacy Tasks. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):1105–1115, January 2025. Conference Name: IEEE Transactions on Visualization and Computer Graphics. doi:10.1109/TVCG.2024.3456155.
- 4 Sibusiso Biyela, Kanta Dihal, Katy Ilonka Gero, Daphne Ippolito, Filippo Menczer, Mike S. Schäfer, and Hiromi M. Yokoyama. Generative AI and science communication in the physical sciences. *Nature Reviews Physics*, 6(3):162–165, March 2024. doi:10.1038/s42254-024-00691-7.
- 5 Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: an open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pages 8359–8388, Vienna, Austria, July 2024. JMLR.org.
- 6 Anthony G Cohn and Robert E Blackwell. Evaluating the Ability of Large Language Models to Reason About Cardinal Directions. In Benjamin Adams, Amy L. Griffin, Simon Scheider, and Grant McKenzie, editors, *16th International Conference on Spatial Information Theory (COSIT 2024)*, volume 315 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 28:1–28:9, Dagstuhl, Germany, 2024. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISSN: 1868-8969. doi:10.4230/LIPIcs.COSIT.2024.28.
- 7 Yu Feng, Linfang Ding, and Guohui Xiao. GeoQAMap - Geographic Question Answering with Maps Leveraging LLM and Open Knowledge Base. In Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise, editors, *12th International Conference on Geographic Information Science (GIScience 2023)*, volume 277 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 28:1–28:7, Dagstuhl, Germany, 2023. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISSN: 1868-8969. doi:10.4230/LIPIcs.GIScience.2023.28.
- 8 Amy L. Griffin, , and Anthony C. Robinson. How do people understand maps and will AI ever understand them? *International Journal of Cartography*, 0(0):1–8, 2025. Publisher: Taylor & Francis. doi:10.1080/23729333.2025.2481692.
- 9 Majid Hojati and Rob Feick. Large Language Models: Testing Their Capabilities to Understand and Explain Spatial Concepts. In Benjamin Adams, Amy L. Griffin, Simon Scheider, and Grant McKenzie, editors, *16th International Conference on Spatial Information Theory (COSIT 2024)*, volume 315 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 31:1–31:9, Dagstuhl, Germany, 2024. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISSN: 1868-8969. doi:10.4230/LIPIcs.COSIT.2024.31.

- 10 Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, 43(2):1–55, March 2025. arXiv:2311.05232 [cs]. doi:10.1145/3703155.
- 11 Yuhan Ji and Song Gao. Evaluating the Effectiveness of Large Language Models in Representing Textual Descriptions of Geometry and Spatial Relations. In Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise, editors, *12th International Conference on Geographic Information Science (GIScience 2023)*, volume 277 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 43:1–43:6, Dagstuhl, Germany, 2023. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISSN: 1868-8969. doi:10.4230/LIPIcs.GIScience.2023.43.
- 12 Peter Kedron, Sarah Bardin, Joseph Holler, Joshua Gilman, Bryant Grady, Megan Seeley, Xin Wang, and Wenxin Yang. A Framework for Moving Beyond Computational Reproducibility: Lessons from Three Reproductions of Geographical Analyses of COVID-19. *Geographical Analysis*, 56(1):163–184, 2024. doi:10.1111/gean.12370.
- 13 Rob Kitchin. The practices of mapping. *Cartographica*, 43(3):211–215, September 2008. doi:10.3138/carto.43.3.211.
- 14 Rob Kitchin, Chris Perkins, and Martin Dodge. Thinking about maps. In *Rethinking Maps: New Frontiers in Cartographic Theory*. Routledge, New York, NY, USA, 2009.
- 15 Markus Konkol, Christian Kray, and Max Pfeiffer. Computational reproducibility in geoscientific papers: Insights from a series of studies with geoscientists and a reproduction study. *International Journal of Geographical Information Science*, 33(2):408–429, February 2019. Publisher: Taylor & Francis. doi:10.1080/13658816.2018.1508687.
- 16 Eftychia Koukouraki and Christian Kray. Map Reproducibility in Geoscientific Publications: An Exploratory Study. In Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise, editors, *12th International Conference on Geographic Information Science (GIScience 2023)*, volume 277 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 6:1–6:16, Dagstuhl, Germany, 2023. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISSN: 1868-8969. doi:10.4230/LIPIcs.GIScience.2023.6.
- 17 Eftychia Koukouraki and Christian Kray. A systematic approach for assessing the importance of visual differences in reproduced maps. *Cartography and Geographic Information Science*, 0(0):1–16, 2024. Publisher: Taylor & Francis. doi:10.1080/15230406.2024.2409920.
- 18 National Academies of Sciences, Engineering, and Medicine. *Reproducibility and Replicability in Science*. National Academies Press, Washington, D.C., September 2019. Pages: 25303. doi:10.17226/25303.
- 19 Frank O. Ostermann, Daniel Nüst, Carlos Granell, Barbara Hofer, and Markus Konkol. Reproducible Research and GIScience: An Evaluation Using GIScience Conference Papers. In Krzysztof Janowicz and Judith A. Verstegen, editors, *11th International Conference on Geographic Information Science (GIScience 2021) - Part II*, volume 208 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 2:1–2:16, Dagstuhl, Germany, 2021. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISSN: 1868-8969. doi:10.4230/LIPIcs.GIScience.2021.II.2.
- 20 Scheider Simon, Jim Jones, Alber Ipia, and Carsten Keßler. Encoding and Querying Historic Map Content. In Joaquín Huerta, Sven Schade, and Carlos Granell, editors, *Connecting a Digital Europe Through Location and Place*, 2014. doi:10.1007/978-3-319-03611-3_15.
- 21 Edward Tufte. *The visual display of quantitative information*. Cheshire: Graphic Press, 2001.
- 22 Kristýna Vlková, Vladimír Zýka, Cristian Remus Papp, and Dušan Romportl. An ecological network for large carnivores as a key tool for protecting landscape connectivity in the Carpathians. *Journal of Maps*, 20(1):2290858, December 2024. Publisher: Taylor & Francis. doi:10.1080/17445647.2023.2290858.

13:12 Assessing Map Reproducibility with Visual Question-Answering

- 23 Jinwen Xu and Ran Tao. Map Reading and Analysis with GPT-4V(ision). *ISPRS International Journal of Geo-Information*, 13(4):127, April 2024. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute. doi:10.3390/ijgi13040127.
- 24 Lu Ying, Yingcai Wu, and Jean-Daniel Fekete. Exploring the Reproducibility for Visualization Figures in Climate Change Report. In Helen-Nicole Kostis, Mark SubbaRao, Yvonne Jansen, and Robert Soden, editors, *IEEE VIS 2024 Workshop on Visualization for Climate Action and Sustainability*, October 2024. URL: <https://inria.hal.science/hal-04744236>.

Guiding Geospatial Analysis Processes in Dealing with Modifiable Areal Unit Problems

Guoray Cai ✉ 🏠 

College of Information Sciences and Technology, The Pennsylvania State University,
University Park, PA, USA

Yue Hao ✉ 

College of Information Sciences and Technology, The Pennsylvania State University,
University Park, PA, USA

Abstract

Geospatial analysis has been widely applied in different domains for critical decision making. However, the results of spatial analysis are often plagued with uncertainties due to measurement errors, choice of data representations, and unintended transformation artifacts. A well known example of such problems is the *Modifiable Areal Unit Problem* (MAUP) which has well documented effects on the outcome of spatial analysis on area-aggregated data. Existing methods for addressing the effects of MAUP are limited, are technically complex, and are often inaccessible to practitioners. As a result, analysts tend to ignore the effects of MAUP in practice due to lack of expertise, high cognitive loads, and resource limitations. To address these challenges, this paper proposes a machine-guidance approach to augment the analyst's capacity in mitigating the effect of MAUP. Based on an analysis of practical challenges faced by human analysts, we identified multiple opportunities for the machine to guide the analysts by alerting to the rise of MAUP, assessing the impact of MAUP, choosing mitigation methods, and generating visual guidance messages using GIS functions and tools. For each of the opportunities, we characterize the behavior patterns and the underlying guidance strategies that generate the behavior. We illustrate the behavior of machine guidance using a hotspot analysis scenario in the context of crime policing, where MAUP has strong effects on the patterns of crime hotspots. Finally, we describe the computational framework used to build a prototype guidance system and identify a number of research questions to be addressed. We conclude by discussing how the machine guidance approach could be an answer to some of the toughest problems in geospatial analysis.

2012 ACM Subject Classification Information systems → Geographic information systems; Information systems → Spatial-temporal systems; Computing methodologies → Artificial intelligence

Keywords and phrases Machine Guidance, Geo-Spatial Analysis, Modifiable Areal Unit Problem (MAUP)

Digital Object Identifier 10.4230/LIPIcs.GIScience.2025.14

1 Introduction

Geospatial analysis plays a critical role in a range of domains [30]. For example, public health professionals used geospatial analysis to track disease outbreaks and plan interventions. During the COVID-19 pandemic, analysts used GISystems to map infection hotspots, model transmission patterns, and allocate healthcare resources efficiently [38]. Practical applications of geospatial analysis in these professional domains involve complicated processes of managing multiple datasets, selecting appropriate spatial scales and methods for analysis, and interpreting geographic patterns. This can be extremely challenging for people without adequate GIS expertise [54] and spatial thinking skills [31, 36, 34].

Due to the unique nature of geographical data, spatial analysis results often suffer from uncertainties in data accuracies, measurement frameworks, transformation artifacts, and spatial heterogeneity [40]. Addressing these uncertainties is essential for ensuring reliable



© Guoray Cai and Yue Hao;

licensed under Creative Commons License CC-BY 4.0

13th International Conference on Geographic Information Science (GIScience 2025).

Editors: Katarzyna Sila-Nowicka, Antoni Moore, David O'Sullivan, Benjamin Adams, and Mark Gahegan;

Article No. 14; pp. 14:1–14:18



Leibniz International Proceedings in Informatics

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

conclusions and decisions. In particular, the Modifiable Areal Unit Problem (MAUP) [20, 51] is a well-known issue that often makes the results of a spatial analysis unreliable. Although the concept of MAUP and related factors is well documented, most analysts choose to ignore MAUP effects in practice due to lack of expertise, high cognitive loads, and resource constraints [50, 26]. Even if analysts are committed to addressing the effects of MAUP, there is very little help and guidance on how to decide the proper strategies and methods in a specific problem-solving context.

To bridge this skill gap for addressing modifiable area unit problem in spatial analysis, we propose a **machine guidance** approach that captures the knowledge and experience necessary for dealing with MAUP into an intelligence agent. While human analysts conduct spatial analysis, a *machine guidance agent* is capable of monitoring the progression of the spatial analysis process and volunteers help and guide in two ways: (1) detect situations where MAUP takes effect and (2) direct users to take proactively measures to mitigate its impact on analytical results. Designing such a machine guidance agent requires that we answer a number of research questions:

1. *Why do analysts tend to ignore MAUP in spatial analysis?* We identified seven (7) reasons why people failed to address MAUP effectively (see Section 3.3). This analysis provides us insights on opportunities for machine guidance.
2. *What factors contribute to the level of MAUP effects?* The effects of MAUP on analytical results could range from *negligible* to *serious* depending on the degree of spatial autocorrelation and spatial heterogeneity, data aggregation methods, and the choices of scale and area units (see Section 3.1). Understanding these causal factors leads to ideas and methods to mitigate MAUP effects.
3. *What are the methods and tools available to address the effects of MAUP?* We synthesize the scattered literature and identify eight methods that are used to help analysts understand the nature and extent of MAUP effects and minimize the effects on the analysis (see Section 3.2). Using these methods requires a significant level of GIS expertise and is cognitively challenging.
4. *What are the opportunities and strategies of machine guidance in addressing MAUP?* Machine guidance exhibits helpful behavior that should be offered only when MAUP arises and when users need help mitigating the effects of MAUP. We identify seven recognizable opportunities and prescribe guidance strategies for them (see Section 4).
5. *How would users (analysts) experience machine guidance?* We demonstrate how users experience guidance by presenting a scenario of use in the context of crime hotspot analysis where the machine guidance agent helps the analyst in dealing with the MAUP. Through the scenario, we gain insight into the expected behaviors of machine guidance.
6. *How can we enable machine guidance computationally?* We show how machine guidance can be enabled computationally by a software agent that can engage with users in collaborative problem solving. Our computational framework was inspired by guidance research in visual analytics [10, 11], advances in mixed-initiative interfaces [53], and intention-based interactions with GIS [9].

By answering the above research questions systematically, this paper contributes to a theoretical foundation of machine guidance in GIScience research. Developing machine guidance tools for geospatial analysis is our long-term goal, and we provide here an initial framework for exploring the design challenges in both conceptual and computational levels.

2 Machine Guidance Approach to Address the MAUP

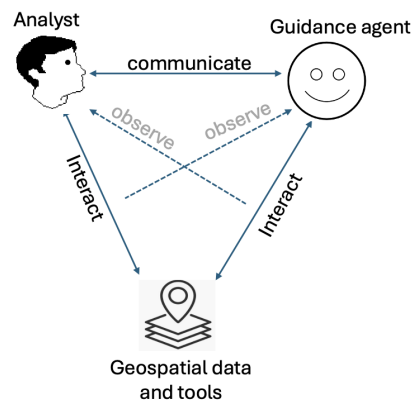


Figure 1 Collaborative Agent architecture of machine guidance.

Machine guidance is an active process of addressing the cognitive challenges and expertise gaps of users that hinder their analytical progress [11]. This approach argues for solving complex and difficult problems by bringing human and computer into a collaborative work relationship [61]. Collaboration is a process in which two or more agents work together to achieve a shared goal. In our case, we introduce a *machine guidance agent* to partner with a *human agent* in spatial analytic activities.

Figure 1 shows the collaborative relationship between human analysts and the guidance agent. A machine guidance agent is an intelligent computational agent that actively assists users during analytical processes by offering contextual guidance, recommendations, and feedback [12]. It can recognize when the analyst encounters difficulties and how to help [11] by integrating reasoning, planning, and communication.

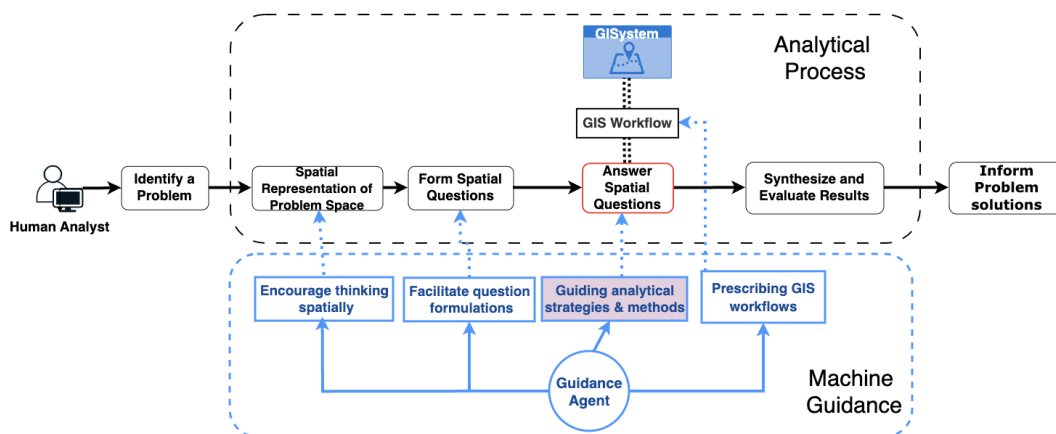


Figure 2 Machine guidance Approach to Supporting Geospatial Analytic Process.

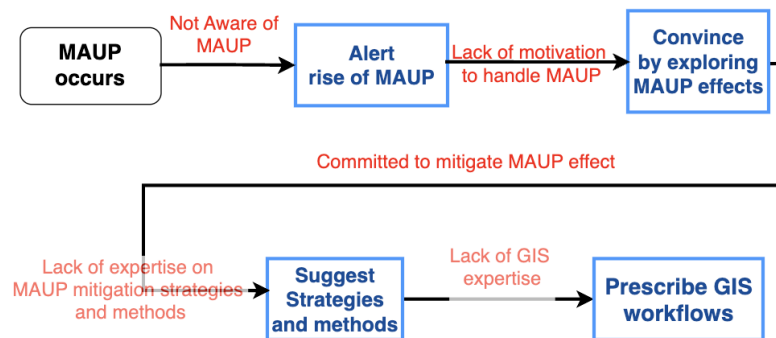
A key capability of a machine guidance agent is to monitor the progression of a spatial analysis process and to volunteer help and guide when needed. As illustrated in Figure 2, the process of solving a domain-specific problem using geospatial analysis generally starts with developing a spatial representation of the problem, followed by the formulation of spatial

questions and the assembly of analytical GIS workflows. Throughout this process, a machine guidance agent works alongside to assert necessary guidance when the human analyst gets lost in navigating the problem and solution space and to steer users away from any dangers and risks under uncertainties.

The task of guiding analysts in dealing with MAUP effects is the responsibility of the box labeled **Guiding Analytical Strategies and Methods**. Zooming into this box, our current work aims at the following two objectives (also summarized in Figure 3):

Obj-1 Building awareness of MAUP effects. The guidance agent actively monitors the analytical process to identify indicators of MAUP occurrences (such as the use of area-aggregated data for analysis). When an MAUP issue occurs, the guidance agent should alert its dangers and potential effects. If the analyst ignores it or is reluctant to address it, the guidance agent plays a role in convincing the analyst to do more exploration to understand the effects on the analytical conclusions.

Obj-2 Bridging the gaps of expertise in mitigating MAUP effects. If the analyst is committed to addressing the MAUP effects, the guidance agent will direct or assist the process of experimenting with multiple spatial units and scales, applying various methods to verify and confirm the choices of area units, and prescribing GIS workflows for proper implementation. Machine guidance simplifies this process by automating repetitive tasks, providing statistical references, and offering immediate feedback on potential solutions. This allows analysts to focus on steering the analysis to achieve confident results.



■ **Figure 3** Machine guidance objectives in dealing with MAUP.

Given the above objectives, it is important to establish a deep understanding of how MAUP arises in spatial analysis, what factors contribute to the serenity of MAUP effects, and what methods and tools are available to explore and mitigate MAUP effects. We will answer the above questions through synthesizing the literature.

3 Nature of the Modifiable Areal Unit Problem (MAUP)

Many applications of geospatial analysis use area-aggregated data as the primary unit of analysis [64, 33]. Data aggregation by area units smooths out local variations, potentially masking important spatial patterns and heterogeneity within areal units. Spatial analysis using area-aggregated data often relies on the assumption of internal uniformity within each area unit. This assumption is rarely true in real world contexts, where factors such as population density, land use, and environmental conditions can vary considerably within a single region. A key issue stems from the wide variety of potential spatial units available

for analysis, including administrative boundaries, census tracts, natural zones, and regular grids. The results of spatial analyses can differ markedly depending on which of these areal units is selected. Openshaw [49] demonstrated this phenomenon by showing how correlation coefficients changed when smaller spatial units were aggregated into larger ones. His findings revealed that correlation values can fluctuate between different spatial scales. This effect, known as the Modifiable Areal Unit Problem (MAUP), undermines the credibility of analyses based on arbitrarily chosen spatial units, casting doubt on the reliability and validity of the resulting conclusions.

The effects of MAUP on analytical results could range from *negligible* to *very serious*. This variability of MAUP effects can be explored by comparing analytical results on different spatial scales (thus the *scale effect* [27]) or using different zoning schemes (thus *zoning effects* [18]). Fotheringham and Wong [25] demonstrated that spatial aggregation introduces biases that vary depending on the chosen scale. This highlights how the choice of spatial scale significantly impacts analytical outcomes, emphasizing the importance of selecting an appropriate scale for an analysis. The *zoning effect*, on the other hand, arises from the specific configuration of spatial units. Even with the same number of zones, different boundary arrangements can produce drastically different statistical outcomes. Openshaw and Taylor [50] experimented with the use of alternative configurations of counties to compute the strength of correlation and they showed that the results of correlation coefficients ranging from 0.265 to 0.862, highlighting the inherent instability in spatial analysis.

3.1 Factors that Cause MAUP Effects

Although MAUP is a general concern in spatial analysis, the actual effect of MAUP on the validity of spatial analysis results could be negligible in some cases and highly problematic in other cases. It is very important to understand the key factors that contribute to the magnitudes of MAUP effects. Here, we synthesize the literature and highlight four major factors.

- F1 The Nature of Boundaries of Area Units.** The boundaries of area units could be *functional* (e.g., natural regions, watersheds, transportation zones) or arbitrary (for example, grids, hexagon). Spatial analysis should avoid arbitrary delineated area boundaries and align with natural boundaries when possible [63]. For example, in crime mapping, the use of square grids can cut through natural neighborhoods, distorting patterns. Instead, mapping crime hotspots using police districts or neighborhoods tends to generate more reliable results.
- F2 Data Aggregation Methods.** Data aggregation methods, such as summing, averaging, or interpolation, determine how data values are combined within spatial units. Different aggregation methods affect both the scale effect (how the results change with different levels of aggregation) and the zoning effect (how results change with different boundary configurations) [37]. The choice of data aggregation methods directly influences the representation and interpretation of spatial patterns, adding another layer of complexity to the MAUP.
- F3 The Degree of Spatial Autocorrelation.** Spatial autocorrelation reflects the similarity between nearby observations. When strong positive spatial autocorrelation is present, neighboring areas tend to have similar values. Aggregating them into larger units inflates spatial dependence, potentially exaggerating trends [41]. The size of areal units significantly influences the strength of spatial autocorrelation, with larger units generally exhibiting lower levels of autocorrelation compared to smaller ones [14]. If

data is aggregated into arbitrary zones, highly autocorrelated spatial data can produce misleading results, as patterns depend on the aggregation scheme rather than the underlying process [47].

- F4 The Scales and Complexities of Spatial Processes.** Spatial processes are mechanisms that generate observable patterns. Examples include natural processes (e.g. erosion and climate change) [19, 52] or human-driven processes (such as migration and urban expansion) [4, 3]. These processes shape spatial patterns across geographical spaces [17]. Since each process operates in a certain scale, the pattern they generate is likely to be in similar scale. Therefore, choosing area units for analysis should consider the alignment with the scales of the underlying processes of the observed patterns [24]. If the chosen area unit in a spatial analysis is inconsistent with the scale of the processes, the effect of MAUP would be worse. What complicates the above rule is that the patterns of real-world phenomena may be the result of multiple processes at different scales interacting in space [65]. This could make any choice of area units seem arbitrary [25].
- F5 Impact of Spatial Heterogeneity.** Spatial heterogeneity refers to the variation in spatial patterns, relationships, and statistical properties at different locations in a study area. This implies that the processes governing spatial phenomena do not operate uniformly across space, leading to location-dependent variations in data distributions and relationships. Spatial heterogeneity violates the *stationarity* assumptions by many statistical models, such as Ordinary Least Squares (OLS) regression, which assume that the relationships between variables are constant across space. The degree of spatial heterogeneity can change depending on the spatial scale or level of aggregation. Aggregating data into larger units (e.g., counties versus census tracts) may mask local variations and distort spatial patterns, which could lead to larger MAUP effects [37].

3.2 Methods for Addressing MAUP

Methods for addressing the MAUP target its underlying causes identified in the last section. Some of the methods (such as sensitivity analysis and multi-scale analysis) help analysts to understand the extent of MAUP effects. Other methods help to choose appropriate area units to mitigate the effects of MAUP by tackling the causal factors of MAUP (as listed in Section 3.1). We discuss a few commonly used methods and their contexts of use.

- M1 Multi-Scale Analysis** conducts analyses at multiple spatial scales. A multi-scale analysis typically begins with small-scale spatial units and then aggregates to larger units as necessary. This strategy ensures that event concentrations at both micro and macro levels are captured, aligning with the analytical context and addressing practical limitations such as data availability and collection challenges [5]. For example, Jelinski [37] used this method to assess how changes in spatial resolution from census tracts to counties affect statistical results.
- M2 Sensitivity Analysis.** Sensitivity analysis runs the same analysis at multiple times by systematically varying the boundary configurations (e.g., administrative zones vs. equal-area grids vs. hexagons) of area units to test the stability of results [49]. For example, voting analysis may be repeated on changing district boundaries to see if electoral outcomes remain stable under different zoning schemes. The method can help to draw the analyst's attention to the serenity of MAUP effects [50].
- M3 Fitness of Use.** Instead of seeking a single “best” unit, analysts should consider the *fitness for use* as the principle when choosing area units for aggregation. For example, analyzing crime hotspots for policing decisions should consider what spatial zones used for deciding police dispatching decisions. If police ward precincts areas are used for

dispatching police, then, analysis should use ward precincts areas if possible. The condition is that the choice is adequate for fulfilling the analytical objectives in a given context [42, 15].

- M4 Respect Scales and Boundaries of Spatial Processes.** Based on our understanding of the relationship between the nature of spatial processes and MAUP effects (F4), the choice of area unit and aggregation scale should reflect the properties of the underlying processes that created the patterns in the data [24]. Because application domains are concerned with different phenomena and different analytical goals, the choice of spatial units is likely to be domain-specific and goal-specific. If we know that a process is operating at a particular scale, then, the choice of spatial units for analysis should respect that scale. Similarly, if the process underlying a pattern create certain boundary conditions, the choice of area unit boundaries for analysis should also respect the this property to minimize the effect of MAUP due to (F1). For example, Buzzelli [7] used census data to study the correlation of patterns between residents of chinese origin and indian origin and he hinted on the need for interpretive skills of a human geographer to draw insights from residential segregation processes.
- M5 Spatial Smoothing Techniques.** Spatial smoothing techniques help mitigate the effects of the Modifiable Areal Unit Problem (MAUP) by reducing abrupt variations caused by arbitrary spatial unit definitions. For example, Kelsall and Wakefield [39] used kernel density estimation to create continuous surfaces from discrete areal data. Spatial interpolation techniques (e.g., Kriging and Inverse Distance Weighting) predicts values at unsampled locations, reducing dependency on arbitrary zone definitions. This method is to used to mitigate the effect of MAUP due to (F3).
- M6 Measuring Spatial Non-Stationarity and Local Variations.** To address the impact of spatial heterogeneity to MAUP effect, measures of spatial non-stationarity and Local Variations, such as Geographically Weighted Regression (GWR) [6], Local Moran's I [2], and Getis-Ord Gi [28], provide insight on the level of local variations. This insight could help the analysts to choose spatial units for analysis to reduce the impact of MAUP.
- M7 Exploratory Spatial Data Analysis (ESDA) techniques.** ESDA methods can be used to detect and mitigate MAUP effects by evaluating spatial patterns at multiple scales and aggregations. For example, by computing and visualizing *Moran's I* [2] for different aggregation levels, analysts can get a sense if spatial autocorrelation remains stable across scales. If stable, the results are less affected by MAUP. If *Moran's I* fluctuates, it suggests strong MAUP effects. ESDA techniques provides insights into the spatial structure and helps identify appropriate scales for analysis. Visualization methods can be used to compare and analyze differences and variations in results [50, 25].

3.3 Practical Challenges of Addressing Modifiable Areal Unit Problems

Despite the rich set of methods to understand and mitigate the MAUP effect (as reviewed in Section 3.2), the effect of MAUP in practical spatial analysis is often overlooked, ignored, or not adequately addressed [25, 18]. This behavior can be explained by understanding the challenges faced by human analysts when dealing with MAUP effects. Here, we discuss seven (7) challenges that explain why people fail to address MAUP effectively.

- C1 Lack of Awareness.** Human analysts keep their attention on answering analytical questions [31]. They may not be aware at the time when an MAUP issue arises. When a stage of spatial analysis involves the use of area-aggregated data, an analyst may not understand how MAUP can affect their analysis. This happens to people even if they have learned MAUP in geography and GIS courses [16, 47].

- C2 Perceived Insignificance of MAUP.** Even when analysts are fully aware of the presence of MAUP-related issues in their analysis, they may choose to ignore them, believing that the impact is too minor to justify the effort required to address it. This belief was partially established by prior research findings. For example, Openshaw [50] showed that the effects of MAUP are often subtle and context-dependent, making it easy to dismiss its importance. Dark and Bram [18] found that the analyst often hold a wrong belief that the conclusions drawn on one scale or zoning scheme will hold on for the other, although this is rarely the case. This has led some analysts to choose not to act on MAUP issues.
- C3 Data Availability.** Exploring the effect of MAUP on spatial analytic outcome requires the availability of data at different scales of area aggregation and different zoning schemes. In reality, data are often available only at specific administrative or aggregated levels (e.g., census tracts, districts), limiting the ability to analyze at finer resolutions. High-resolution data and individual-level data can be difficult to obtain [20, 25, 60]. Wong [64] noted that researchers frequently rely on preaggregated data due to privacy concerns, cost, or logistical constraints, which restricts their ability to address MAUP.
- C4 Practical Constraints.** Applying GIS methods (as discussed in section 3.2) to mitigate the MAUP effects costs time, computing resources, and human effort. In real world practices, analysts are often under pressure to deliver actionable results and have limited time and resources, making it impractical for analysts to fully explore how scale or zoning choices influence results [18].
- C5 Convenience of Choice on Default Spatial Units.** Analysts often use default spatial units (e.g. administrative boundaries) for convenience without considering their appropriateness for the analysis. Dark and Bram [18] argue that administrative boundaries are often arbitrary and may not be aligned with the underlying spatial processes being studied.
- C6 Lack of Expertise in Applying Complex Methods.** As noted in Section 3.2, addressing MAUP requires a thorough understanding of the available methods, the conditions under which specific methods can be applied, and how to implement them using matched tools in a GIS. The expertise in choosing and applying the appropriate methods to practical problems is rarely available to most analysts.
- C7 Lack of Tool Support.** Methods for mitigating MAUP effects are challenging to practice because they require support from GIS tools. Although relevant analytical tools are available in popular GIS systems, such as ArcGIS, they are not structured and streamlined for the purpose of dealing with MAUP effects. The application-dependent nature of MAUP effects makes it difficult to design tool support.

Machine guidance can help human analysts overcome each of the above challenges to achieve reliable and confident analytical results. Machine guidance can monitor the spatial analytical process and alert analysts when the MAUP effect comes into play (C1), convince them by showing them the danger of not addressing MAUP (C2, C3, C4, C5), and provide suggestions on proper methods and tools to mitigate MAUP effects (C6, C7).

4 When and How to Guide?

Given the inherent complexities and challenges of addressing MAUP, there are critical moments where machine guidance can effectively assist analysts. In this section, we use the seven key challenges in addressing MAUP (as outlined in Section 3.3) to pinpoint critical moments when guidance is needed. Table 1 characterizes the possible guidance *opportunities*

corresponding to the seven user challenges. For example, guidance can be inserted when the system detects that the analysis involves the use of area-aggregated data in geostatistical analysis (G1).

■ **Table 1** Opportunities For Asserting Machine Guidance.

[Note: C1–C7 correspond to the user challenges described in Section 3.3. S1–S11 are guidance strategies described in Table 2.]

User Challenges	Opportunities
C1: The analyst is unaware of the MAUP.	G1: The guidance should inform the analyst that the MAUP effects can be involved (S1) and thus encourage the analyst to explore more on its effects (S2, S3).
C2: The analyst does not know whether MAUP is critical in the current situation.	G2: The guidance assesses whether there is a significant effect of the MAUP. If yes, the analyst will be convinced to address the MAUP by showing what are the possible consequences if MAUP is not addressed (S2, S3).
C3: Limited data availability for exploring and mitigating the MAUP effects.	G3: The guidance can help the analyst by 1) looking for other data sets that are disaggregated and can be applied in the context (S4), and by 2) directing the analyst to consider other data processing and modeling methods (S5, S6).
C4: The analyst has limited time and resources.	G4: The guidance could recommend suitable methods that are less time consuming for the analyst to pursue (S4, S8, S9). The guidance could take initiative to generate results of multi-scale analysis and present them visually as an effort to alert and convince the analyst (S1, S2).
C5: The MAUP is not addressed due to convenience of use.	G5: The guidance examines whether applied units are appropriate by considering: 1) whether units are aligned with the spatial process in a given context (S9), 2) how much effects are involved based on statistical variations (S6), 3) simulating and comparing results using other units (S7).
C6: The analyst has trouble applying suitable methods to address the MAUP.	G6: Guidance can help the analyst determine which methods are helpful at the moment and automate the processing steps to reduce the complexities (S6, S7).
C7: The analyst has difficulties implementing suitable methods with GIS tools.	G7: Guiding the analyst by recommending proper GIS workflows tools to use (S11). If the analyst has a preference but does not know how to perform it, the guidance will assist the translation of workflows into GIS procedures for a particular platform (S3, S6).

To take advantage of the guidance opportunities identified in Table 1, the guidance agent must form intention to volunteer guidance and formulate a strategy to generate guidance messages. Table 2 describes the guidance strategies we use as design rationales for our guidance agent. For each strategy, we specify the goals that can be achieved and prescribe a recipe for action. These guidance strategies are consistent with the guidance objectives described in Figure 3. It is important to note that the guidance agent does not dictate how the analyst deals with the MAUP issue. If the agent believes that the effect of MAUP should be handled, the guidance agent will convince the analyst to do more explorations and suggest suitable methods and operations to mitigate the MAUP effects according to the prescribed action recipes.

■ **Table 2** Detailed strategies of the machine guidance. Each strategy can be applied to act on one or more guidance opportunities in Table 1.

Strategy	Goal	Recipe for Action	Required Machine Knowledge
S1	Alert the existence of MAUP	Use maps to show variations when different spatial units are used (M8).	Data availability Aggregation and visualization methods
S2	Convince the analyst to explore more	Visualize the patterns of the spatial phenomenon and interacted factors (M8).	Data availability Knowledge on domain and phenomenon
S3	Convince the analyst to explore more	Calculate and use statistical indicators to measure and inform possible effects (M8).	Data availability Aggregation and statistical methods
S4	Mitigate the MAUP effects	Use disaggregated data instead (M5).	Data availability
S5	Mitigate the MAUP effects	Use smoothing techniques (M6).	Data availability Spatial smoothing methods
S6	Mitigate the MAUP effects	Use spatial models and statistics to consider the local variations (M7).	Data availability Spatial modeling methods
S7	Recommend suitable units	Recommend suitable units by considering the statistical variations when using different units (M1, M2).	Analytical methods
S8	Recommend suitable units	Choose units that are aligned with the analytical goal (M3).	Contextual knowledge Knowledge on domain and phenomenon
S9	Recommend suitable units	Choose units that are aligned with the spatial process (M4).	Data availability Knowledge on domain and phenomenon Spatial process
S10	Recommend suitable units	Measure the local variations to find suitable units (M7, M8).	Aggregation and statistical methods
S11	Recommend suitable tools and workflow	Recommend suitable tools or workflow in addressing the MAUP.	Data availability Knowledge on domain and phenomenon Analytical methods

It is important to emphasize that the set of strategies prescribed in Table 2 is a significant finding of this paper. It fills a knowledge gap between mitigation goals (Table 1) and GIS methods (described in Section 3.2). Despite the abundance of methods available to address MAUP, there has been little understanding of how to effectively match and apply these methods to specific mitigation goals. For example, multi-scale analysis (M1) and sensitivity analysis (M2) are frequently cited as methods useful for dealing with MAUP, but exactly how to apply them is a knowledge inaccessible to most analysts.

5 How Users Experience Machine Guidance?

To illustrate how a human analyst experiences interacting with the guidance system, we present a hypothetical scenario in which a public safety analyst uses geospatial analysis of crime hotspots to inform police actions.

Danny, a public safety analyst at the Baltimore City Police Department, is responsible for planning crime prevention strategies. He is charged with developing a police patrol plan on how to dispatch officers to neighborhoods based on crime hotspot patterns. Since the department has a limited number of police force to dispatch, it must ensure that the dispatch plan generates a measurable reduction in crime rates. It is very important that Danny derives reliable and trustworthy results from his analysis. He has access to ArcGIS Desktop and crime data from the last few months.

Danny is familiar with basic concepts and methods of GIS analysis, but he is not an expertise in GIS tools and algorithms. Danny is representative of a class of analysts who are experts in their fields but have limited or no knowledge of geospatial analysis methods and tools [48, 62]. These analysts lack specialized training in GIScience or have only surface knowledge of MAUP.

Danny has access to a crime incident dataset that contains ten types of crime (see the picture of MG 1 in Table 3). Each type of crime has different underlying mechanisms and processes that produce the crime patterns. Criminogenic situations can vary in scale, duration and impact, affecting entire regions or specific groups [23]. This raises challenges with respect to the selection of an appropriate spatial unit to identify hotspot areas [44].

Based on the narrative of the scenario above, we present a hypothetical sequence of interactions between the User (Danny) and the guidance agent (MG) in Table 3. This hypothetical dialogue showcases how machine guidance can systematically address MAUP by raising awareness, recommending alternative methods, and providing statistical support to improve decision making under practical constraints.

As Danny analyzes crime hotspots in neighborhoods, the guidance agent actively monitors the process and detects potential MAUP effects due to spatial aggregation. At step (User 3), the system sensed that the user is unaware of the MAUP problem, an alert guidance is initiated to warn the user about the danger. To help the user understand the risk of MAUP effects, the guidance agent repeated the analysis using an alternative spatial unit, census tracts, and showed the user that the result is significantly different (MG 3-a). To further convince the user to take steps to mitigate risk, the agent computed LISA (as a prescribed strategy S8), suggesting strong MAUP effects (MG 3-b). These maps and messages convinced Danny to commit serious effort to mitigate the MAUP effect.

At step (User 4), Danny acknowledges the rise of a MAUP issue and decides to explore the likely effects. However, Danny does not know how to proceed. He asks for help directly and the guidance agent suggests using disaggregated data where available (S4) and offers

■ **Table 3** A hypothetical interaction session of guiding users in addressing MAUP effect.

Dialogue	Guidance Visual Messages
<p>User 1: Let us take a look if crime patterns in Baltimore City have changed recently.</p> <p>MG 1: We found 4,983 crime incidents in Baltimore City happened in the last two months. The map (MG 1) shows ten types of crime.</p> <p>User 2: Our police officers typically divide their responsibilities by neighborhoods. Let us find out where are the crime hot spots by neighborhoods.</p> <p>MG 2: As show on the map (MG 2), crime hotspots are noted by red circles. They are derived from crime density patterns by neighborhoods.</p> <p>User 3: Ok, that is what I need. I will proceed on using this hot spot patterns for police dispatch.</p> <p>MG 3: Wait! I would like to remind you that the pattern shown on MG 2 may not be reliable! Let me explain. If you aggregate crimes by census tracts (instead of neighborhoods), the hotspot pattern will look like the one in MG 3-a, which is significantly different. This is a well known problem in spatial analysis on area-aggregated data. MG 3-b shows strong local spatial dependencies, which indicates that your analytical results could be very unstable and could be misleading.</p> <p>User 4: This is so important to know! How can I make sure my results are reliable?</p> <p>MG 4: Let me help you to explore more. One way to find out if the hotspots on MG 2 are trustful is view crime density as continuous surface. MG 4 shows a Kernal Density pattern where A lighter color indicates higher density and vice versa. You may compare MG 2 with this density surface and make adjustment of hotspot regions.</p> <p>User 5: Thanks! I was able to determine those hotspots with good confidence. However, there are too many neighborhoods in the hotspot regions, and I need to prioritize those neighborhoods that are most troubled.</p> <p>MG 5: Sure. I can further gauge the strength of those hotspots using a measure called Getis-Ord Gi. The map (MG 5) shows the Gi measure for each neighborhood, indicating their relative degree of confidence as a hot spot.</p>	<p>The visual messages consist of six maps labeled MG 1 through MG 5, each illustrating a different geospatial analysis technique for crime data in Baltimore City.</p> <ul style="list-style-type: none"> MG 1: Crime Incidents in Baltimore City - A map showing the distribution of 4,983 crime incidents across the city, categorized by type: HOMICIDE, BURGLARY, SHOOTING, ROBBERY, RAPE, LARCENY, and AGG. ASSAULT. It also shows census tracts. MG 2: Density by Neighborhood - A map showing crime density by neighborhood, with hotspots indicated by red circles. The legend indicates density levels: 0 - Mean, Mean - 2 Mean, 2 Mean - 3 Mean, 3 Mean - 4 Mean, and > 4 Mean. MG 3-a: Density by Census Tract - A map showing crime density by census tract, illustrating how the pattern changes when aggregated to a larger spatial unit. MG 3-b: Indicator of Autocorrelation by Neighborhood - A map showing Moran's I values for neighborhoods, indicating local spatial dependencies. The legend indicates Moran's I < 0, 0, and Moran's I > 0. MG 4: Kernel Density - A map showing a continuous kernel density surface of crime incidents, where lighter colors indicate higher density. MG 5: Hot Spots based on Gi* - A map showing hot spots based on the Getis-Ord Gi* measure, indicating the relative degree of confidence as a hot spot. The legend indicates Hot Spot - 90% Confidence, Hot Spot - 80% Confidence, Hot Spot - 70% Confidence, Hot Spot - 60% Confidence, Hot Spot - 50% Confidence, Hot Spot - 40% Confidence, Hot Spot - 30% Confidence, Hot Spot - 20% Confidence, Hot Spot - 10% Confidence, and Not Significant.

KDE (S5) as an alternative method for density calculations, mitigating the distortions introduced by arbitrary spatial units. In this stage, Danny was guided to choose mitigation methods. He was also assisted in executing a proper GIS workflow for exploratory analysis. For practical reasons, Danny is not free to choose any area units other than neighborhood boundaries. The guidance agent adapted a strategy to verify the hotspots using kernel

density surface representation (MG 4). MG 5 was generated using the Incremental Spatial Autocorrelation tool ¹ to determine an appropriate spatial scale (M8), which is then applied as the distance banding parameter for Hotspot Analysis ² with Gi* statistics (M8). Such statistical validation is used as additional evidence to convince Danny that he should take measures to minimize uncertainty and improve the reliability of their conclusions.

6 Computational Framework of Machine Guidance

Our approach would not be complete without discussing the feasibility of achieving our design goals through machine intelligence. To demonstrate the feasibility of machine guidance, we are developing a prototype design that supports the guidance behavior demonstrated in the scenario of Table 3. A full discussion on that prototype implementation is beyond the scope of this paper. However, we do want to briefly describe the computational frameworks employed and shed light on the practicality of implementing machine guidance.

6.1 An Agent-based Computational Framework

Our implementation of a guidance agent is primarily based on the SharedPlan model of human-computer collaboration [35, 53]. This model is capable of representing the intentional structures of agent collaborations and reasoning for planning future actions under uncertainty. This adaptability is crucial in guiding geospatial analysis, where problem-solving evolves dynamically with new information.

Our guidance agent is a specialized type of collaborative interface agent [46]. The guidance agent is able to communicate and observe the actions of the human analyst and vice versa. A crucial part of successful collaboration is knowing when a particular analytical action has been performed and what are the intended analytical goals. SharedPlan model has been successfully applied in geo-analytical tasks, helping GISystems infer user intent beyond direct commands and reducing ambiguity through dialogue-based interactions [8]. Cai [9] showed that the analytical intentions of the analyst can be recognized with certain domain knowledge. Using the SharePlan model in a conversational agent, basic GIS analysis tasks can be done through conversations with the interface agent. Our work extends this agent framework for mixed-initiative guidance.

Another source of inspiration is research on guidance in the field of visual analytics [11, 12, 10, 55]. Guidance was defined as a computational system that actively assists users during analytical processes by offering contextual guidance, recommendations, and feedback [11, 12]. Machine guidance identifies when help is needed and determines the type of assistance to provide [11] by integrating reasoning, planning, and domain knowledge. Recent works such as Lotse [58] and AdViCE [29] bridge theoretical concepts with practical applications and allow analysts to receive better assistance in data exploration and visualization tasks. However, designing guidance systems that scale across different data domains and user expertise levels remains a significant challenge [22]. Practical applications to support geospatial analysis remain limited, despite similar challenges, such as the need to make critical decisions while lacking the expertise and tools.

¹ <https://desktop.arcgis.com/en/arcmap/latest/tools/spatial-statistics-toolbox/incremental-spatial-autocorrelation.htm>

² <https://desktop.arcgis.com/en/arcmap/latest/tools/spatial-statistics-toolbox/hot-spot-analysis.htm>

6.2 Knowledge Representation and Reasoning

Design of guidance agents must answer a number of questions: (1) What is knowledge and expertise represented? (2) What reasoning abilities are needed? (3) What kinds of sensing skills are needed to monitor changes in contexts? (4) What communication behaviors are expected? These questions can be partially answered by observing the communication and interaction patterns in the scenario presented in Table 3.

- *The system must actively monitor the analytical process, identify the current analysis stage, detect whether MAUP is involved, and recognize when the analyst encounters difficulties.* This requires the system to have a sensing capability and be able to keep track of the analytical process to determine when help is needed and what form of guidance should be provided.
- *Guidance should not merely follow the analyst's actions but must take the initiative to intervene when necessary.* This requires that the system must be able to form intention to act based on reasoning about what is helpful to do for the user.
- *It is important to convince the analyst to address the MAUP effect before suggesting mitigation methods and strategies.* Thus, the system must be able to plan complex actions based on reasoning about strategies, methods, and tools.
- *The system must be adaptive and context aware, tailoring guidance based on specific analytical domains, available data, and the analytical goals of the analyst.* This involves dynamically inferring the analyst's intentions, understanding the current analytical context, and determining how to deliver relevant guidance.

7 Discussion and Conclusion

The Modifiable Areal Unit Problem (MAUP) continues to pose a significant challenge in GIScience, yet discussions surrounding its causes, consequences, and solutions remain fragmented. Although existing research has primarily emphasized the scientific implications of MAUP, practical strategies for addressing it in real-world applications are still limited and underdeveloped [50, 25]. Our analysis reveals that many analysts tend to overlook MAUP or underestimate its impact, underscoring a critical disconnect between theoretical understanding and practical implementation.

Our work contributes to a practical approach to address MAUP in geospatial analysis. We proposed to introduce an intelligent agent to guide analysts in mitigating the effect of MAUP. As the first step toward this long-term goal, this paper established a preliminary theory of machine guidance by answering a number of fundamental research questions. We identified multiple opportunities for the machine to guide the analysts by alerting to the rise of MAUP, assessing the impact of MAUP, choosing mitigation methods, and generating visual guidance messages using GIS functions and tools. In terms of choosing what guidance features to be designed, we set two sets of objectives machine guidance in MAUP: (1) building awareness (2) supplement user's expertise in mitigating MAUP effects. This level of understanding allows for further refinement and formalization of the related expertise in computational systems.

MAUP in geospatial analysis poses challenges in identifying its causes, selecting mitigation strategies, and interpreting scale-dependent results [63, 50]. Machine guidance has the potential to provide a proactive solution for addressing MAUP by alerting analysts to potential consequences, offering suitable methods, and facilitating executions in the GISystem. Given the resolution-dependent nature of geographic data [32], the selection of appropriate methods is crucial. Visual guidance, such as standardized map comparisons (Table 3), helps analysts

interpret MAUP effects more effectively [57, 13], reducing the likelihood of overlooking its impact [21, 59, 1]. Addressing MAUP through machine guidance demonstrates its potential to enhance geospatial analysis in various domains by expanding its knowledge base and integrating domain-specific solutions [45, 43, 56].

The work presented in this paper is the first step towards the goal of active machine guidance when analysts encounter MAUP during geospatial analysis. Although we made a convincing argument for the feasibility of machine guidance and its capacity to address MAUP, the scientific merit of this approach needs to be assessed by the usefulness of the tool (machine guidance agent) when it is implemented, refined, and tested. Our ongoing research focuses on evaluating and refining the proposed strategies to ensure practical applicability. Observing how participants interact with the system, our aim is to gain a deeper understanding of *when and how the guidance should be introduced* when addressing the MAUP. We are collecting data on user experience and feedback and identify areas for improvement. We apply a human-centered approach to further refine both the conceptual and computational components. The findings of the study of machine guidance are likely to inspire and inform researchers in both GIS and Human-Computer Interaction (HCI) regarding the design of interactive components in GISystems.

References

- 1 Natalia Andrienko and Gennady Andrienko. *Exploratory analysis of spatial and temporal data: a systematic approach*. Springer Science & Business Media, 2006.
- 2 Luc Anselin. Local indicators of spatial association—lisa. *Geographical analysis*, 27(2):93–115, 1995.
- 3 Luc Anselin. *Spatial econometrics: methods and models*, volume 4. Springer Science & Business Media, 2013.
- 4 Michael Batty. *Cities and complexity: understanding cities with cellular automata, agent-based models, and fractals*. The MIT press, 2007.
- 5 Patricia L Brantingham, Paul J Brantingham, Mona Vajihollahi, and Kathryn Wuschke. Crime analysis at multiple scales of aggregation: A topological approach. *Putting crime in its place: Units of analysis in geographic criminology*, pages 87–107, 2009.
- 6 Chris Brunsdon, A. Stewart Fotheringham, and Martin E. Charlton. Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geographical Analysis*, 28(4):281–298, October 1996.
- 7 Michael Buzzelli. Modifiable areal unit problem. *International encyclopedia of human geography*, page 169, 2019.
- 8 Guoray Cai, Hongmei Wang, Alan M MacEachren, and Sven Fuhrmann. Natural conversational interfaces to geospatial databases. *Transactions in GIS*, 9(2):199–221, 2005. doi:10.1111/J.1467-9671.2005.00213.X.
- 9 Guoray Cai, Bo Yu, and Dong Chen. Modeling and communicating the conceptual intent of geo-analytical tasks for human-gis interaction. *Transactions in GIS*, 17(3):353–368, 2013. doi:10.1111/TGIS.12040.
- 10 Davide Ceneda, Natalia Andrienko, Gennady Andrienko, Theresia Gschwandtner, Silvia Miksch, Nikolaus Piccolotto, Tobias Schreck, Marc Streit, Josef Suschnigg, and Christian Tominski. Guide me in analysis: A framework for guidance designers. In *Computer Graphics Forum*, volume 39(6), pages 269–288. Wiley Online Library, 2020. doi:10.1111/CGF.14017.
- 11 Davide Ceneda, Theresia Gschwandtner, Thorsten May, Silvia Miksch, Hans-Jörg Schulz, Marc Streit, and Christian Tominski. Characterizing guidance in visual analytics. *IEEE transactions on visualization and computer graphics*, 23(1):111–120, 2016. doi:10.1109/TVCG.2016.2598468.

- 12 Davide Ceneda, Theresia Gschwandtner, Thorsten May, Silvia Miksch, Marc Streit, and Christian Tominski. Guidance or no guidance? a decision tree can help. In *Euro VA@ Euro Vis*, pages 19–23, 2018. doi:10.2312/EUROVA.20181107.
- 13 Spencer Chainey, Svein Reid, and Neil Stuart. When is a hotspot a hotspot? a procedure for creating statistically robust hotspot maps of crime. *Innovations in GIS*, 9:21–36, 2002.
- 14 Andrew David Cliff and J Keith Ord. Spatial processes: models & applications. (*No Title*), 1981.
- 15 Alexis Comber and Paul Harris. The importance of scale and the maup for robust ecosystem service evaluations and landscape decisions. *Land*, 11(3):399, 2022.
- 16 National Research Council, Life Studies, Board on Earth Sciences, Geographical Sciences Committee, Committee on Support for Thinking Spatially, and The Incorporation of Geographic Information Science Across the K-12 Curriculum. *Learning to think spatially*. National Academies Press, 2005.
- 17 Noel Cressie. *Statistics for spatial data*. John Wiley & Sons, 2015.
- 18 Shawna J Dark and Danielle Bram. The modifiable areal unit problem (maup) in physical geography. *Progress in Physical Geography*, 31(5):471–479, 2007.
- 19 William E Dietrich, Dino G Bellugi, Leonard S Sklar, Jonathan D Stock, Arjun M Heimsath, and Joshua J Roering. Geomorphic transport laws for predicting landscape form and dynamics. *Geophysical Monograph-American Geophysical Union*, 135:103–132, 2003.
- 20 Jennifer L Dungan, JN Perry, MRT Dale, Pousty Legendre, S Citron-Pousty, M-J Fortin, A Jakomulska, M Miriti, and MS2002 Rosenberg. A balanced view of scale in spatial statistical analysis. *Ecography*, 25(5):626–640, 2002.
- 21 Karin Eberhard. The effects of visualization on judgment and decision-making: a systematic literature review. *Management Review Quarterly*, 73(1):167–214, 2023.
- 22 Alex Endert, William Ribarsky, Cagatay Turkay, BL William Wong, Ian Nabney, I Díaz Blanco, and Fabrice Rossi. The state of the art in integrating machine learning into visual analytics. In *Computer Graphics Forum*, volume 36(8), pages 458–486. Wiley Online Library, 2017. doi:10.1111/CGF.13092.
- 23 Ihor Fedchak, Oleksandr Kondratiuk, Anatolii Movchan, and Svyatoslav Poliak. Theoretical foundations of hot spots policing and crime mapping features. *Social and Legal Studies*, 1(7):174–183, 2024.
- 24 A Stewart Fotheringham and Mehak Sachdeva. Scale and local modeling: new perspectives on the modifiable areal unit problem and simpson’s paradox. *Journal of Geographical Systems*, 24(3):475–499, 2022. doi:10.1007/S10109-021-00371-5.
- 25 A Stewart Fotheringham and David WS Wong. The modifiable areal unit problem in multi-variate statistical analysis. *Environment and planning A*, 23(7):1025–1044, 1991.
- 26 Andrew U Frank. Qualitative spatial reasoning: Cardinal directions as an example. *International journal of geographical information science*, 10(3):269–290, 1996. doi:10.1080/02693799608902079.
- 27 Charles E Gehlke and Katherine Biehl. Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association*, 29(185A):169–170, 1934.
- 28 Arthur Getis and J Keith Ord. The analysis of spatial association by use of distance statistics. *Geographical analysis*, 24(3):189–206, 1992.
- 29 Oscar Gomez, Steffen Holter, Jun Yuan, and Enrico Bertini. Advice: Aggregated visual counterfactual explanations for machine learning model validation. In *2021 IEEE Visualization Conference (VIS)*, pages 31–35. IEEE, 2021. doi:10.1109/VIS49827.2021.9623271.
- 30 MF Goodchild and PA Longley. The future of gis and spatial analysis. *Geographical information systems*, 1:567–580, 1999.
- 31 Michael F Goodchild. Geographical information science. *International journal of geographical information systems*, 6(1):31–45, 1992. doi:10.1080/02693799208901893.
- 32 Michael F Goodchild. Scale in gis: An overview. *Geomorphology*, 130(1-2):5–9, 2011.

- 33 Michael F Goodchild, Luc Anselin, Richard P Appelbaum, and Barbara Herr Harthorn. Toward spatially integrated social science. *International Regional Science Review*, 23(2):139–159, 2000.
- 34 Michael F Goodchild and Robert P Haining. Gis and spatial data analysis: Converging perspectives. *Papers in Regional Science*, 83(1):363–385, 2004.
- 35 Barbara J Grosz and Sarit Kraus. Collaborative plans for complex group action. *Artificial Intelligence*, 86(2):269–357, 1996. doi:10.1016/0004-3702(95)00103-4.
- 36 Robert Haining. Designing spatial data analysis modules for geographical information systems. *Spatial analysis and GIS*, pages 45–64, 1994.
- 37 Dennis E Jelinski and Jianguo Wu. The modifiable areal unit problem and implications for landscape ecology. *Landscape ecology*, 11:129–140, 1996.
- 38 Dayun Kang, Hyunho Choi, Jong-Hun Kim, and Jungsoo Choi. Spatial epidemic dynamics of the covid-19 outbreak in china. *International journal of infectious diseases*, 94:96–102, 2020.
- 39 Julia Kelsall and Jonathan Wakefield. Modeling Spatial Variation in Disease Risk: A Geostatistical Approach. *Journal of the American Statistical Association*, 97(459):692–701, September 2002.
- 40 Mei-Po Kwan. The uncertain geographic context problem. *Annals of the Association of American Geographers*, 102(5):958–968, 2012.
- 41 Sang-Il Lee, Monghyeon Lee, Yongwan Chun, and Daniel A Griffith. Uncertainty in the effects of the modifiable areal unit problem under different levels of spatial autocorrelation: A simulation study. *International Journal of Geographical Information Science*, 33(6):1135–1154, 2019. doi:10.1080/13658816.2018.1542699.
- 42 Stefan Leyk, Andrea E Gaughan, Susana B Adamo, Alex De Sherbinin, Deborah Balk, Sergio Freire, Amy Rose, Forrest R Stevens, Brian Blankespoor, Charlie Frye, et al. The spatial allocation of population: a review of large-scale gridded population data products and their fitness for use. *Earth System Science Data*, 11(3):1385–1409, 2019.
- 43 Paul A Longley, Michael F Goodchild, David J Maguire, and David W Rhind. *Geographic information science and systems*. John Wiley & Sons, 2015.
- 44 Cehong Luo. The modifiable areal unit problem (maup) in the spatial analysis of crime and socio-economic indicators, 2023.
- 45 Alan M MacEachren and Menno-Jan Kraak. Research challenges in geovisualization. *Cartography and geographic information science*, 28(1):3–12, 2001.
- 46 Pattie Maes. Agents that reduce work and information overload. *Communications of the ACM*, 37(7):30–40, 1994. doi:10.1145/176789.176792.
- 47 David Manley. Scale, aggregation, and the modifiable areal unit problem. In *Handbook of regional science*, pages 1711–1725. Springer, 2021.
- 48 Timothy L Nyerges. Cognitive issues in the evolution of gis user knowledge. In *Cognitive aspects of human-computer interaction for geographic information systems*, pages 61–74. Springer, 1995.
- 49 Stan Openshaw. A million or so correlated coefficients: three experiment on the modifiable areal unit problem. *Statistical applications in the spatial sciences*, 1979.
- 50 Stan Openshaw. Ecological fallacies and the analysis of areal census data. *Environment and planning A*, 16(1):17–31, 1984.
- 51 S Openshaw. A million or so correlation coefficients, three experiments on the modifiable areal unit problem. *Statistical applications in the spatial science*, pages 127–144, 1979.
- 52 Camille Parmesan and Gary Yohe. A globally coherent fingerprint of climate change impacts across natural systems. *nature*, 421(6918):37–42, 2003.
- 53 Charles Rich, Candace L Sidner, and Neal Lesh. Collagen: Applying collaborative discourse theory to human-computer interaction. *AI magazine*, 22(4):15–15, 2001.
- 54 Joacim Rocklöv and Henrik Sjödin. High population densities catalyse the spread of covid-19. *Journal of travel medicine*, 27(3):taaa038, 2020.

- 55 Floarea Serban, Joaquin Vanschoren, Jörg-Uwe Kietz, and Abraham Bernstein. A survey of intelligent assistants for data analysis. *ACM Computing Surveys (CSUR)*, 45(3):1–35, 2013. doi:10.1145/2480741.2480748.
- 56 Terry A Slocum, Connie Blok, Bin Jiang, Alexandra Koussoulakou, Daniel R Montello, Sven Fuhrmann, and Nicholas R Hedley. Cognitive and usability issues in geovisualization. *Cartography and geographic information science*, 28(1):61–75, 2001.
- 57 Terry A Slocum, Robert B McMaster, Fritz C Kessler, and Hugh H Howard. *Thematic cartography and geovisualization*. CRC Press, 2022.
- 58 Fabian Sperrle, Davide Ceneda, and Mennatallah El-Assady. Lotse: A practical framework for guidance in visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):1124–1134, 2022. doi:10.1109/TVCG.2022.3209393.
- 59 John Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive science*, 12(2):257–285, 1988. doi:10.1207/S15516709C0G1202_4.
- 60 Andrew Swift, Lin Liu, and James Uber. Maup sensitivity analysis of ecological bias in health studies. *GeoJournal*, 79:137–153, 2014.
- 61 Loren G Terveen. Overview of human-computer collaboration. *Knowledge-Based Systems*, 8(2-3):67–81, 1995. doi:10.1016/0950-7051(95)98369-H.
- 62 Carol Traynor. Putting power in the hands of end users: a study of programming by demonstration, with an application to geographical information systems. In *CHI 98 conference summary on Human factors in computing systems*, pages 68–69, 1998. doi:10.1145/286498.286533.
- 63 David Wong. The modifiable areal unit problem (maup). In *The SAGE Handbook of Spatial Analysis*, pages 105–123. SAGE Publications, Ltd, London, 2009.
- 64 D.W. Wong. Modifiable areal unit problem. In Rob Kitchin and Nigel Thrift, editors, *International Encyclopedia of Human Geography*, pages 169–174. Elsevier, Oxford, 2009.
- 65 Jianguo Wu, K Bruce Jones, and Orie L Loucks. *Scaling and uncertainty analysis in ecology*. Springer, 2006.

Accommodating Space-Time Scaling Issues in GAM-Based Varying Coefficient Models

Alexis Comber¹  

School of Geography, University of Leeds, UK

Paul Harris  

Sustainable Agriculture Sciences, Rothamsted Research, Harpenden, UK

Chris Brunsdon  

National Centre for Geocomputation, National University of Ireland, Maynooth, Ireland

Abstract

The paper describes modifications to spatial and temporal varying coefficient (STVC) modelling, using Generalized Additive Models (GAMs). Previous work developed tools using Gaussian Process (GP) thin plate splines parameterised with location and time variables, and has presented a space-time toolkit in the `stgam` R package, providing wrapper functions to the `mgcv` R package. However, whilst thin plate smooths with GP bases are acceptable for working with spatial problems they are not for working with space *and* time combined. A more robust approach is to use a tensor product smooth with GP basis. However, these in turn require correlation function length scale or range parameters (ρ) to be defined. These are distances (in space or time) at which the correlation function falls below some value, and can be used to indicate the scale of spatial and temporal dependencies between response and predictor variables (similar to geographically weighted bandwidths). The paper describes the problem in detail, illustrates an approach for optimising ρ and methods for determining model specification.

2012 ACM Subject Classification Mathematics of computing → Regression analysis; Information systems → Spatial-temporal systems; Mathematics of computing → Spline models

Keywords and phrases Spatial Analysis, Spatiotemporal Analysis

Digital Object Identifier 10.4230/LIPIcs.GIScience.2025.15

Funding This research was supported by UKRI (ESRC) funding ES/Y006259/1 under the Digital Footprints scheme.

1 Introduction

Previous research has described the use of Generalized Additive Models (GAMs) [13, 12] with Gaussian Process (GP) smooths as an approach for spatially varying coefficient (SVC) modelling [5, 6]. The proposed geographical GP-GAM has been shown to have all of the advantages of the SVC brand leader, geographically weighted regression (GWR) [2] and its multiscale variant (MGWR) [18, 15, 11] in modelling and capturing any spatial dependencies between the target variable and individual predictor variables (hence *multiscale*), and none of the disadvantages: GWR-based approaches are subject to local collinearity, they generate a collection of local models rather than a single one, until recently MGWR was only specified for Gaussian responses, MGWR is unable to support out of sample predictions and all GWR-based approaches re-use individual observations in multiple local models. Ideas extending the use of GAMs with GPs for SVC model construction into space-time analyses for *spatial and temporally* varying coefficient (STVC) models have been proposed [4] and, at the time of writing, are currently under review [7]. In parallel the `stgam` package [8] was developed to

¹ Corresponding author



© Alexis Comber, Paul Harris, and Chris Brunsdon;
licensed under Creative Commons License CC-BY 4.0

13th International Conference on Geographic Information Science (GIScience 2025).

Editors: Katarzyna Sila-Nowicka, Antoni Moore, David O'Sullivan, Benjamin Adams, and Mark Gahegan;
Article No. 15; pp. 15:1–15:9



Leibniz International Proceedings in Informatics

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

provide a framework for informed SVC and STVC model construction, through functions that wrapped around the `mgcv` packages `gam()` function [17] to fit a range of spatial, temporal, and spatiotemporal varying coefficient models, to investigate the nature of any space-time dependencies present in the data and to inform SVC and STVC model specification.

However, we have subsequently become aware of a number of problems with simply extending the `stgam` approach to SVC modelling to STVC models. This paper describes these, some potential solutions and their potential implementation in a revised `stgam` package.

2 Background

2.1 The original big idea

Increasingly the space-time data we analyse and use are *not* collected under some grand over-arching experimental design, nor for the purposes we intend to use them for. As such, the big idea behind `stgam` workflows is that it is naive to construct models that make assumptions about the presence and nature of data space-time dependencies, whether for the purposes of prediction or inference (process understanding). Instead these need to be explicitly examined and the most appropriate model form determined. This position is in contrast to a classic statistics perspective where data are considered to be a realisation of carefully considered data collection activity, constructed in such a way as to allow specific hypothesis to be tested.

In this context, most widely used approaches for capturing spatial and temporal dependencies in data and process heterogeneity are flawed because they assume the presence space-time interactions and dependencies. Examples include the alignment of lagged responses to nearby lagged variables in autoregressive moving average models, and existing GAM-based approaches that consider variable interaction over space but with only the aim of model selection and penalization and not process inference [10].

GAMs can be specified with smooths or splines to model non-linear relationships. These are constructed from basis functions that can include single or multiple predictor variables. If a predictor variable is included in a smooth with geographic location (X and Y) then non-linear relationships over space can be modelled - an SVC model. If the smooth is specified with geographic location and time of observation (X , Y and T) then an STVC model is specified. To illustrate this, consider each predictor variable in a SVC or STVC model. It can be specified in 3 different ways:

- i. It is excluded from the model.
- ii. It is included in the SVC / STVC model as a standard parametric response (as in an OLS regression).
- iii. It is included in the SVC / STVC model in a smooth with location (X and Y) but not time.

There are a further 3 ways that each covariate can be specified in an STVC model:

- iv. It is included in the STVC model in a smooth with time (T) but not location.
- v. It is included in the STVC model in a smooth with location and time (X , Y and T).
- vi. It is included in the STVC model in 2 separate smooths, one with location (X and Y) and the other with time (T).

The intercept is treated in a similar way, but without it being absent. Thus for any SVC with k predictor variables there are $2 \times 3^{k+1}$ potential models and for any STVC there are $5 \times 6^{k+1}$ potential models.

2.2 The `stgam` R package

The `stgam` package [8] was created to provide a wrapper around GAM functionality in the `mgcv` package [17], and to allow the user to investigate the different ways each predictor variable could be specified within the GP smooth (as described above), and thus model selection and specification. Its workflow determines the most plausible model given the data and it includes functions that i) create multiple SVC and STVC models defined in different ways as described above, ii) determine the probability of each being the correct model given the data, iii) combine multiple plausible models and, iv) generate spatially and temporally varying coefficient estimates.

`stgam` is underpinned by 2 core concepts. First, the need to test for the presence and nature of any spatial and temporal dependences and thus to determine whether to include each predictor variable in the model, including within GAM smooths, thereby informing SVC / STVC model specification. In `stgam` and related initial work [3], this is done by creating multiple models with each predictor variable specified in different ways, as described above. The probability of each model being the best model given the data is then determined and the most probable model is selected. The second core concept in `stgam` is to evaluate the probability of each potential model being the correct model given the data. This is determined from the model Bayesian Information Criterion (BIC) value [16]. It has been shown elsewhere that the marginal posterior probability of observing the data D given the model M_i – $\Pr(D|M_i)$ – can be approximated using BIC [3], and used to derive the probability of any individual model M_i being the correct model given the data. Of course this is under the assumption that the true model, although unknown, is one of the potential models being evaluated. If multiple models are highly probable, then model averaging can be performed with model weighting from the BIC-derived posterior beliefs in each model.

However, since the release of the first version of the `stgam` package and paper submission to conferences and peer reviewed journals, some issues with the approach to STVC model construction have been identified. These derive from the way that the approach for SVC modelling GAM with GP smooth was extended to STVC models and the use of GP basis from the `mgcv` package. Whilst `mgcv` GP bases may be appropriate for spatial problems and SVC model construction, they are not for STVC models. This is because the `mgcv` GP basis results in spatial models with a potentially suboptimal estimate of the length scale of the correlation basis functions used in the GP basis, and thus temporal models that are erroneous for all time series except those with the longest of length scales.

2.3 The problem in a bit more detail

The substantive problem with the `mgcv` GP basis is that it only fits within the penalized spline class of models that can be estimated by `mgcv` if the basis function parameters are specified. The critical parameter here is the correlation function length scale or range parameter, ρ which may be specified as spherical, power exponential, or as one of three forms of Matérn correlation with κ equal to 1.5, 2.5, or 3.5. The ρ parameter determines the distance at which the correlation function falls below some value. If ρ is not supplied, then `mgcv` fits a smooth using Matérn correlation functions with $\kappa = 1.5$ and $\rho = \max_{i,j} \|\mathbf{x}_i - \mathbf{x}_j\|$ as the basis functions, for any pair of points x_i and x_j , following [14]. If an order penalty is specified then one of the correlation functions is implemented but again using Kammann and Wand's recommendation for the length scale, the maximum distance attained by any pair of points x_i and x_j .

Evidently for SVCs (and similar spatial analyses) this is unlikely to be problematic except in the presence of long or short spatial dependencies. However, for temporal and spatiotemporal data, this specification of length scale in this way **is** problematic: it implies that the correlation between pairs of points will only fall below some small value when those points are separated by an amount of time equivalent to the time series itself. A space-time GP smooth specified in this way will be isotropic, with similar levels of non-linearity in all dimensions, such that the spatial heterogeneity is equal to temporal heterogeneity. This is clearly not correct but the current implementation of `stgam` has no way to control this, because the functions that create different models hardcode a GP basis in the smooths, and there is currently no option to specify an order penalty or ρ . The result is GP smooths specified with pure Kammann and Wand basis functions, despite their unsuitability for time series or space-time problems.

3 Methods

3.1 Proposed improvements

The proposed high level solutions to these problems are 1) to allow users to specify the length scales of the different components, and 2) to fit the spatio-temporal terms via a tensor product smooth of a marginal spatial smooth and a marginal time smooth, both specified with a GP basis, and effectively allowing space-time to be treated in a three-dimensional way. This can be done by specifying a 2D spatial GP smooth for the first margin of the tensor product, and a 1D temporal GP smooth for the second margin. These two bases will smoothly interact creating the desired spatio-temporal varying term due to the tensor product construction.

In `mgcv` syntax, the change in how the GP smooths for each predictor variable `var` are specified as follows:

```
# FROM: a mgcv smooth with GP basis
s(x, y, t, bs = "gp", by = var)
# TO: a mgcv tensor product smooth with GPs
te(x, y, t, d = c(2, 1), bs = rep("gp", 2), by = var)
```

The tensor product smooth requires the marginal basis dimensions to be specified (`d` above). Here the space-time smooth contains 3 covariates composed of a tensor product of a 2-D thin plate regression spline basis for location, and 1-D basis for time.

However, the correlation function length scale or range parameter, ρ also needs to be specified as, both in terms of its form (spherical, power exponential, etc) and distance. In the `mgcv` GAM implementation this is done for each of the margins through the `m` argument passed to the tensor product smooth. This expands the specification of the tensor product smooth with GPs to:

```
# TO: a mgcv tensor product smooth with GPs
# WITH: user specified scale form & length
te(x, y, t, d = c(2, 1), bs = rep("gp", 2), by = var,
    m = list(c(3, rho1), c(3, rho2)))
```

Here `rho1` is the length range for the spatial marginal smooth and `rho2` is that for the temporal marginal. Both are specified with a power of 3, indicating the distance decay of the range. For the spatial margin intervals of 100km for a case study in a large country (USA, China, Brazil, etc) could be explored, 10 km for a small country (UK, Germany, Vietnam, etc) or 1km for a local one. These may be the equivalent of predictor variable bandwidths

in MGWR. For the temporal dimension, intervals of 1, 2, 3 etc years could be explored, depending on how time was recorded in the data. The revised **stgam** package will support investigation of different forms and length scales for spatial and temporal margins.

The ability to vary the scales and lengths ranges in the tensor smooths as above allows each predictor variable to be specified in a way that treats implicitly space-time as three-dimensional. Therefore a second set of investigations will be supported in the revised **stgam** package to allow users to determine which of the possible six forms described above is the appropriate way to include each predictor variable in SVC and STVC models. For example an alternative to the **mgcv** tensor product smooth with GPs, and illustrated above is one that has separate 2-D spatial smooths and 1-D temporal ones:

```
te(x, y, d = 2, bs = "gp", by = var, m = c(3, rho1)) +
  te(t, d = 1, bs = "gp", by = var, m = c(3, rho2))
```

Essentially, this replaces the investigation of different model forms using **mgcv** smooths specified (**s()** in **mgcv**), with ones specified with **mgcv** tensor product smooths (**te()** in **mgcv**). The analysis undertaken in this short paper describes initial work investigating how both of these considerations (space-time lengths and predictor variable form) can be optimised and potentially included in a revised **stgam** package.

3.2 Data and model

The **stgam** package includes two datasets describing annual economic productivity for the 48 contiguous US states (with Washington DC merged into Maryland), 1970 to 1985 from the **plm** R package [9] and the spatial dataset of the 48 contiguous US states **spData** package [1]. The productivity data contains variables describing Private Capital stock (**privC**), Unemployment % (**unemp**) and Public capital investment (**pubC**). The Unemployment variable over time is shown in Figure 1, with 1986 not plotted for aesthetic reasons. The aim of the analysis was

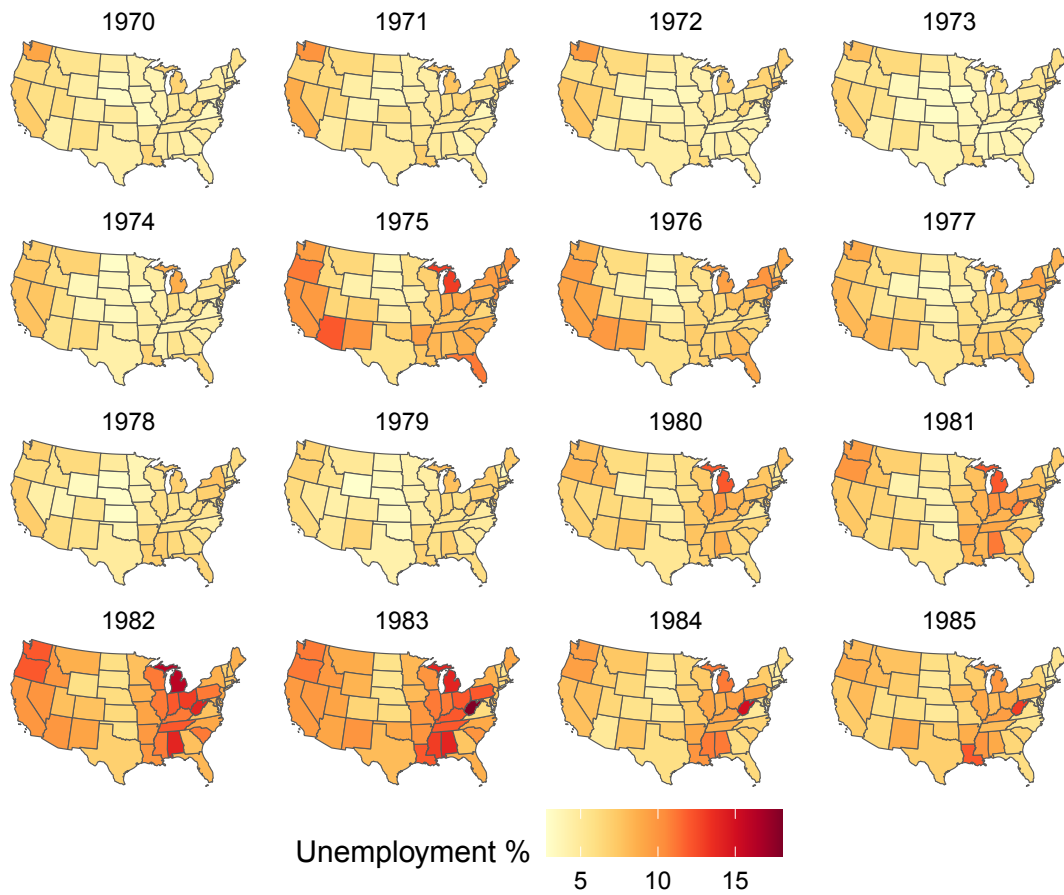
1. to determine optimal ρ values for the correlation function length scales, ρ , a 2D spatial margin and a 1D temporal margin (ρ_s and ρ_t , **rho1** and **rho2** respectively in the above code snippet).
2. to then determine the most appropriate STVC model of Private Capital stock (**privC**), with Unemployment and Public capital investment as predictor variables, specify the tensor product smooths in different ways as described in Section 2.1.

In both cases, ρ and model form were evaluated from the model BIC.

4 Results

The optimal ρ values were determined by creating multiple GAM models with tensor product smooths with GP bases, each with different ρ_s , for the 2D spatial margin and ρ_t for the 1D temporal margin. After investigation of the distances and time series lengths in the data, values for ρ_s from 0 km to 4,500 km in steps of 250 km were explored with values for ρ_t from 0 to 17 years in 1 year steps. A total of 42 models were created with different combinations of ρ_s and ρ_t , and the BIC value for each model extracted. The top ρ values are shown in Table 1 and indicate a ρ_s of 250 km and ρ_t of 9 years. What is interesting is that 250 km is the optimal spatial length range in all of the top 10 models and there is some similar consistency in the temporal length range.

These values for ρ_s and ρ_t specify the length ranges, could be plugged into to a model with each parameter specified in the manner indicated in Section 3.1. However this still assumes that the predictor variables exhibit simultaneous dependencies space-time dependencies with



■ **Figure 1** The % unemployed over US States, 1970-1985.

■ **Table 1** The values of rho1 and rho2 that resulted in the best 10 models.

rho1	rho2	BIC
250	9	5074.882
250	10	5074.893
250	7	5075.093
250	8	5075.252
250	11	5075.675
250	12	5076.220
250	13	5076.938
250	6	5077.692
250	14	5077.719
250	15	5078.704

the target variable (option *v.* in the list in Section 2.1 - *in a smooth with location and time* (X, Y and T)). The second stage in the analysis evaluated different model forms, with each predictor variable specified in one of 6 ways, and the intercept in one of five ways (i.e. 180 models). The models were evaluated using BIC from which the probabilities of the model being the best model were determined in the approach outlined in [3]. The results are

shown in Table 2 and indicate that there is a 4.5% chance that the second ranked model is better than the first, suggesting that the top ranked model can be used. This omits the unemployment predictor variable (**unemp**) and specifies space-time tensor product smooths. It is possible to extract the spatially and temporally varying coefficient estimates from these and to examine the nature of the spatial and temporal dependencies in the data. This is left for future research.

■ **Table 2** The 10 best models, and how the predictor variables were specified within the model, where “—” indicates the absence of a predictor, “Fixed” that a parametric form was specified, “te_S” a spatial tensor product smooth, “te_T” a temporal tensor product smooth and “s_ST” a spatio-temporal tensor product smooth.

Rank	Intercept	unemp	pubC	BIC	Pr(M)
1	te_ST	—	te_ST	4920.984	—
2	te_ST	Fixed	te_ST	4927.090	0.045
3	te_ST	te_T	te_ST	4942.752	0.000
4	te_S	—	te_ST	4963.511	0.000
5	te_S	Fixed	te_ST	4971.211	0.000
6	te_ST	te_S	te_ST	4972.547	0.000
7	te_T + te_S	—	te_ST	4975.040	0.000
8	te_ST	te_T + te_S	te_ST	4976.802	0.000
9	te_T + te_S	Fixed	te_ST	4981.094	0.000
10	te_S	te_T	te_ST	4982.354	0.000

5 Discussion

This short paper describes the next stage in the evolution of an approach for varying coefficient modelling based on GAMs with smooths. It unpicks work described in some published papers [5, 6, 7] for undertaking spatially and temporally varying coefficient modelling, wrapping functionality from the **mgcv** package [17]. Initial work developed spatially varying coefficient (SVC) models using Gaussian Process (GP) smooths that included observation locations, and this framework was extended to include observation location and time. The focus of this extension into space-time modelling was to determine the most appropriate way to specify space-time interactions (dependencies) in the smooths, for example in a single combined space-time smooth or in separate ones for space and for time. However, the type of the smooth is also important. Whilst GP smooths are appropriate for location, they are not for space-time interactions due to assumptions. This is because of the default way that the GP basis correlation function length scale or range parameters (ρ (the distance at which the correlation function falls below some value)) are determined in the **mgcv** GP basis if they are not specified.

This paper details a revised approach to STVC modelling using GAMs with tensor product smooths. Combining a 2D marginal spatial smooth and a 1D marginal time smooth each specified with a GP basis, within tensor product smooth is a more robust way to treat space-time interactions and dependencies. The correlation function length scale, ρ , still needs to be specified. The first part of the analysis optimally determined ρ for both the 2D spatial margin and the 1D temporal margin. In this illustrative case study, these were found to be 250km and 9 years. The second part of the analysis sought to determine the most appropriate model form as described in Section 2.1, with using tensor product smooths and

the optimally determined ρ values. Here the model with the greatest probability of being correct was one with a combined tensor product smooth for the Intercept and for the `pubC` (public capital) predictor variable, and that omitted unemployment from the model.

The determination of the optimal ρ values and the probability of each model being the correct model given the data, both used the model BIC value as described in [7, 3]. However this is not without controversy, as there is a potential lack of a theoretical foundation for the use of BIC for penalized spline models of the form fitted by `mgcv`. For future work, it will be important to establish the viability of a BIC approach to model selection and ρ determination. A second issue is that for many researchers undertaking model selection based on optimising some fit, parsimony or error measure is itself fallacious. The counter argument is that users should just fit the most complex model that they believe is valid for their task in hand. If their understanding of the process being examined is that it is a spatio-temporally varying process then a model specifying that interaction should be fitted. However a counter argument is that increasingly researchers and analysts are working with secondary data, collected for a different purpose and actually part of their job is to determine what kind of spatial, temporal and space-time dependencies are present in the data. Finally, the optimisation of ρ for both the 2D spatial margin and the 1D temporal margin in the tensor product smooths is exciting as this indicates the nature of scales of interactions between the predictor variables and the target variable in the same way that bandwidths do in geographically weighted approaches. Future work will explore how these can be optimised for each predictor variable in a similar way to multiscale geographically weighted approaches and of course their interaction with model selection.

References

- 1 Roger Bivand, Jakub Nowosad, and Robin Lovelace. *spData: Datasets for Spatial Analysis*, 2024. R package version 0.3. URL: <https://CRAN.R-project.org/package=spData>.
- 2 Chris Brunsdon, A Stewart Fotheringham, and Martin E Charlton. Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis*, 28(4):281–298, 1996.
- 3 Chris Brunsdon, Paul Harris, and Alexis Comber. Smarter than your average model-bayesian model averaging as a spatial analysis tool (short paper). In *12th International Conference on Geographic Information Science (GIScience 2023)*. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023. doi:10.4230/LIPIcs.GIScience.2023.17.
- 4 Alexis Comber, Paul Harris, and Chris Brunsdon. Multiscale spatially and temporally varying coefficient modelling using a geographic and temporal gaussian process gam (gtgp-gam). In *Proceedings of 12th International Conference on Geographic Information Science (GIScience 2023)*, volume 277, page 22. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023.
- 5 Alexis Comber, Paul Harris, and Chris Brunsdon. Multiscale spatially varying coefficient modelling using a geographical gaussian process gam. *International Journal of Geographical Information Science*, 38(1):27–47, 2024. doi:10.1080/13658816.2023.2270285.
- 6 Alexis Comber, Paul Harris, Daisuke Murakami, Tomoki Nakaya, Narumasa Tsutsumida, Takahiro Yoshida, and Chris Brunsdon. Encapsulating spatially varying relationships with a generalized additive model. *ISPRS International Journal of Geo-Information*, 13(12):459, 2024. doi:10.3390/IJGI13120459.
- 7 Alexis Comber, Paul Harris, Naru Tsutsumida, Jennie Gray, and Chris Brunsdon. Specifying spatially and temporally varying coefficient models using gams with gaussian process splines. *Transactions in GIS*, submitted.
- 8 Lex Comber, Paul Harris, and Chris Brunsdon. *stgam: Spatially and Temporally Varying Coefficient Models Using Generalized Additive Models*, 2024. R package version 0.0.1.2. URL: <https://CRAN.R-project.org/package=stgam>.

- 9 Yves Croissant, Giovanni Millo, and Kevin Tappe. *plm: Linear Models for Panel Data*, 2025. R package version 2.6-5. URL: <https://CRAN.R-project.org/package=plm>.
- 10 Jakob A Dambon, Fabio Sigrist, and Reinhard Furrer. Joint variable selection of both fixed and random effects for gaussian process-based spatially varying coefficient models. *International Journal of Geographical Information Science*, 36(12):2525–2548, 2022. doi: 10.1080/13658816.2022.2097684.
- 11 A Stewart Fotheringham, Wenbai Yang, and Wei Kang. Multiscale geographically weighted regression (mgwr). *Annals of the American Association of Geographers*, 107(6):1247–1265, 2017.
- 12 T Hastie and R Tibshirani. Generalized additive models. chapman hall & crc. *Monographs on Statistics & Applied Probability. Chapman and Hall/CRC*, 1, 1990.
- 13 Trevor Hastie and Robert Tibshirani. Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398):371–386, 1987.
- 14 EE Kammann and Matthew P Wand. Geoadditive models. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 52(1):1–18, 2003.
- 15 Binbin Lu, Chris Brunsdon, Martin Charlton, and Paul Harris. Geographically weighted regression with parameter-specific distance metrics. *International Journal of Geographical Information Science*, 31(5):982–998, 2017. doi:10.1080/13658816.2016.1263731.
- 16 Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.
- 17 Simon Wood. *mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation*, 2023. R package version 1.9-1. URL: <https://CRAN.R-project.org/package=mgcv>.
- 18 Wenbai Yang. *An extension of geographically weighted regression with flexible bandwidths*. PhD thesis, University of St Andrews, 2014.

What, When, and Where Do You Mean?

Detecting Spatio-Temporal Concept Drift in Scientific Texts

Meilin Shi¹ 

Department of Geography and Regional Research, University of Vienna, Austria

Krzysztof Janowicz 

Department of Geography and Regional Research, University of Vienna, Austria

Zilong Liu 

Department of Geography and Regional Research, University of Vienna, Austria

Mina Karimi 

Department of Geography and Regional Research, University of Vienna, Austria

Ivan Majic 

Department of Geography and Regional Research, University of Vienna, Austria

Alexandra Fortacz 

Department of Geography and Regional Research, University of Vienna, Austria

Abstract

Inundated by the rapidly expanding AI research nowadays, the research community requires more effective research data management than ever. A key challenge lies in the evolving nature of concepts embedded in the growing body of research publications. As concepts evolve over time (e.g., keywords like *global warming* become more commonly referred to as *climate change*), past research may become harder to find and interpret in a modern context. This phenomenon, known as *concept drift*, affects how research topics and keywords are understood, categorized, and retrieved. Beyond temporal drift, such variations also occur across geographic space, reflecting differences in local policies, research priorities, and so forth. In this work, we introduce the notion of *spatio-temporal concept drift* to capture how concepts in scientific texts evolve across both space *and* time. Using a scientometric dataset in geographic information science, we detect how research keywords drifted across countries and years using word embeddings. By detecting spatio-temporal concept drift, we can better align archival research and bridge regional differences, ensuring scientific knowledge remains findable and interoperable within evolving research landscapes.

2012 ACM Subject Classification Information systems → Digital libraries and archives; Computing methodologies → Information extraction; Information systems → Ontologies

Keywords and phrases Concept Drift, Ontology, Large Language Models, Research Data Management

Digital Object Identifier 10.4230/LIPIcs.GIScience.2025.16

Supplementary Material *Software (Source Code and Data):*

<https://github.com/meilinshi/Spatio-temporal-Concept-Drift-in-Scientific-Texts>

archived at `swb:1:dir:928946b53e1ebade1bc2cd2ddfe8a0b23a409111`

1 Introduction

The questions of *what*, *when*, and *where* permeate our daily conversations. When scheduling a group meeting, for instance, we agree on the topic of discussion (e.g., a proposal, *what*), the time (e.g., 10 a.m., *when*), and the location (e.g., a café, *where*). In communicating such

¹ Corresponding author



© Meilin Shi, Krzysztof Janowicz, Zilong Liu, Mina Karimi, Ivan Majic, and Alexandra Fortacz; licensed under Creative Commons License CC-BY 4.0

13th International Conference on Geographic Information Science (GIScience 2025).

Editors: Katarzyna Sila-Nowicka, Antoni Moore, David O'Sullivan, Benjamin Adams, and Mark Gahegan;

Article No. 16; pp. 16:1–16:18



Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

information, we implicitly agree on a particular reference system. For time, we have temporal reference systems such as the Gregorian calendar, the yyyy-mm-dd date format, the 24-hour clock, and so on. For space, we have various geodetic datums, such as WGS 84 and NAD 83, as well as known place names we can refer to. For the *what* question, namely the thematic information, we also need a semantic reference system [23]. In this reference system, an ontology can help ensure that, by “proposal”, we are referring to a research proposal rather than a marriage proposal.

When it comes to ontology modeling and engineering, concepts are often represented as static entities [17]. For example, this is common in a foundational ontology (e.g., DOLCE [15]) to ensure a coherent view and interoperability across domains. In the real world, however, concepts are constantly evolving [19], and their meanings can vary across different social contexts and locations, as seen in the evolving sociocultural definitions of *gender* nowadays. Research in the Semantic Web and the broader knowledge representation and reasoning (KRR) communities has focused on *concept drift* to capture the dynamics of evolving concepts. In this respect, previous work in KRR [49, 14, 44, 8] focused mainly on the temporal aspect of a concept, i.e., the changing meaning of a concept over *time*, and overlooked the spatial perspective that often accompanies it.

In geographic information science (GIScience), constructing an ontology that maps geospatial concepts has always been challenging because of their unique spatio-temporal properties [12, 9]. Geospatial concepts, such as **Mountain** and **Forest**, are different from other general concepts because they do not have clearly defined boundaries nor can they be distinguished in bona fide fashion from neighboring concepts, e.g., **Hill** and **Woods** [42, 43]. For instance, the difference between **Lake** and **Pond** can be affected by seasonal water level variations [28]. This would make downstream tasks, e.g., question answering [33], more challenging. Furthermore, the conceptualization of such landscape concepts may also vary and evolve across languages, cultures, and regions [47, 11].

These challenges are not limited to modeling the aforementioned concepts that are vague geographic features. They also extend to research topics and keywords (e.g., *urban planning*, *climate change*), which we see as signifiers of concepts (i.e., mental representations that categorize areas of research [4]). Although many concepts in this regard exhibit concept drift, geospatial concepts are particularly susceptible due to their inherent dependence on physical, environmental, and sociopolitical contexts. To give a concrete example, the definition of *disasters* could vary significantly depending on local environmental conditions, infrastructure, and response systems. What qualifies as a natural *disaster* in one region (e.g., an earthquake with a magnitude of 6.0 in Haiti) may be labeled differently in another region (e.g., the United States) because of differences in local resilience. In comparison, concepts like *the speed of light* exhibit less spatial variability because of their underlying physical principles.

As Kuhn et al. [24] suggested, we should move space and time from merely being in application domains to becoming foundational aspects of ontologies. While the inherent vagueness in geographic features is not fully resolved, ontologies such as the GeoNames ontology² offer structured representations for these features. However, these efforts have yet to cover more abstract geospatial concepts embedded in research, such as those represented by keywords. Same as geographic features, these concepts are dynamic and spatially grounded, yet they are even more susceptible to societal changes and technological advancements (e.g., the coining of the concept *GeoAI* [21]). In this work, we look into these concepts in scientific research with explicit study areas. We propose an approach to quantify their fluidity and context-dependence across space *and* time via word embeddings. Our long-term goal is to

² <https://www.geonames.org/ontology>

develop an ontology that can address the dynamic nature of these concepts in scientific texts. This contributes to the FAIR principles (Findability, Accessibility, Interoperability, and Reusability) [50] by improving retrieval, reuse, and ensuring the long-term relevance of research [38].

The contributions of this work are as follows:

- We introduce spatio-temporal concept drift, which expands the previous notions of concept drift that focused mainly on temporal changes by incorporating both space *and* time.
- We propose a novel approach to detecting spatio-temporal concept drift in scientific texts via word embeddings.
- We demonstrate how accounting for spatio-temporal concept drift enhances the understanding of concepts in scientific texts, improves recall in FAIR-based research data management systems, and lays the groundwork for ontology learning with large language models (LLMs) in dynamic contexts.

The remainder of this paper is structured as follows. Section 2 introduces the theoretical background of our work. Section 3 reviews related work regarding word embeddings, with a focus on their ability to capture and quantify spatio-temporal variations in semantics. We describe our case study in Section 4, where we use a scientometric dataset in the field of GIScience to assess spatio-temporal concept drift. Section 5 presents the results. Section 6 discusses geographic bias within embeddings and future directions in ontology learning with LLMs. Finally, we conclude our work in Section 7.

2 Background

This section provides the theoretical background for the study on concept drift in knowledge representation. Here, we introduce the notion of spatio-temporal concept drift. This notion adds a spatial dimension to existing definitions of concept drift in the literature. In addition, we discuss the broader implications of spatio-temporal concept drift for research data management (RDM).

2.1 Concepts and Their Representation

The notion of *concept* has many different definitions across or even within domains, such as in linguistics, psychology, computer science, and cognitive science [35]. In this work, we adopt the definition by Stock [45] in information science, which defines a concept as a class containing objects that share certain properties.³

Concepts are fundamental units of meaning and serve as the building blocks of ontologies that help structure knowledge, enable reasoning, and facilitate interoperability. In terms of representation, previous work [49, 44] typically characterized a concept by its label (i.e., name), intension (i.e., defining properties), and extension (i.e., instances that fall under it), in the form $(label(C), int(C), ext(C))$ for a concept C . However, this would only apply to concepts that already existed in predefined ontologies. Verkijk et al. [48] proposed to use embedding techniques to derive vector representations of concepts. While their work focuses on knowledge-graph data, they showcased the ability of embedding techniques to capture flexible, context-aware representations of concepts for natural language data as well, in the form $(label(C), context(C))$.

³ It is worth noting that this definition of *concept* is closer to what cognitive scientists would call a *category*.

In this work, we treat keywords in research publications as representations of underlying concepts. Unlike established ontological categories, research keywords are rapidly evolving as science and society change. This makes them particularly relevant for studying spatio-temporal concept drift. Here, we focus on research keywords also with the aim of developing a structured ontology that can capture their changing meanings over time and space. Such an ontology could contribute to RDM by improving metadata organization, literature retrieval, question answering, and knowledge graph construction in scientific databases.

2.2 Spatio-Temporal Concept Drift

Adding a temporal dimension to concept representation accounts for changes in their meaning over time. The study of *concept drift*, as defined by Wang et al. [49], aims to capture these changes in concept meaning over time. For example, the keyword *global warming* was once the dominant term in research publications, referring to the rise in Earth’s temperature. Over time, *climate change* became more widely used to capture broader climate-induced impacts [26] and account for the fact that *warming* is not uniform. To model a concept C with a temporal component, it can be represented in the form $(label_t(C), int_t(C), ext_t(C))$ or $(label_t(C), context_t(C))$ at time t . Extending this notion to a spatio-temporal dimension means that a concept’s meaning may change both over time and space, e.g., at different rates. Here, we define this phenomenon as *spatio-temporal concept drift*. In this case, a concept can be represented as $(label_{t,s}(C), int_{t,s}(C), ext_{t,s}(C))$ or $(label_{t,s}(C), context_{t,s}(C))$ for a concept C at time t and region s .

Figure 1 illustrates how a concept moves in both time and geographic space. More abstractly, this can be thought of as the trajectory of a concept in a space-time prism [35]. The color gradient indicates (*semantic*) *concept similarity* [36, 39, 34, 22], thereby reflecting changes in its thematic dimension over time and space. Take the keywords *global warming* and *climate change* as an illustrative example. Region s_1 may have already adopted the use of *climate change* since time t_1 , while s_2 still has a mixed use of both terms. Note that, here at t_1 , the variation in the thematic dimension between s_1 and s_2 represents spatial variability, which differs from spatio-temporal concept drift, as it captures regional differences without a temporal dimension [25]. Later by t_3 , s_2 adopts the distinct use of *climate change* and *global warming*, aligning its concept representation with s_1 . Over time, as concepts evolve, their meanings may change gradually, showing a concept drift from t_2 to t_3 in s_1 , or change so much that they diverge into two, showing a concept split⁴ in region s_2 . Ultimately, the two concepts may converge into a shared understanding for both regions (indicated by the semantic similarity between C_1 and C'_1 at t_3).

Even with the advent of semantic search [18], which allows for more flexible query interpretation, concept drift remains a challenge, particularly in RDM and other archival systems. If the past and present keywords are not properly linked, search results may still be skewed toward more recent ones, simply because of their pertinence. This would lead to either incomplete retrieval results or misinterpretation of archival documents. Addressing spatio-temporal concept drift helps ensure that evolving knowledge remains accessible and meaningful across different time periods and regions. It could therefore enhance semantic interoperability overall and support the FAIR principles.

⁴ Definitions of *concept drift* and *concept split* are provided in our earlier work [40].

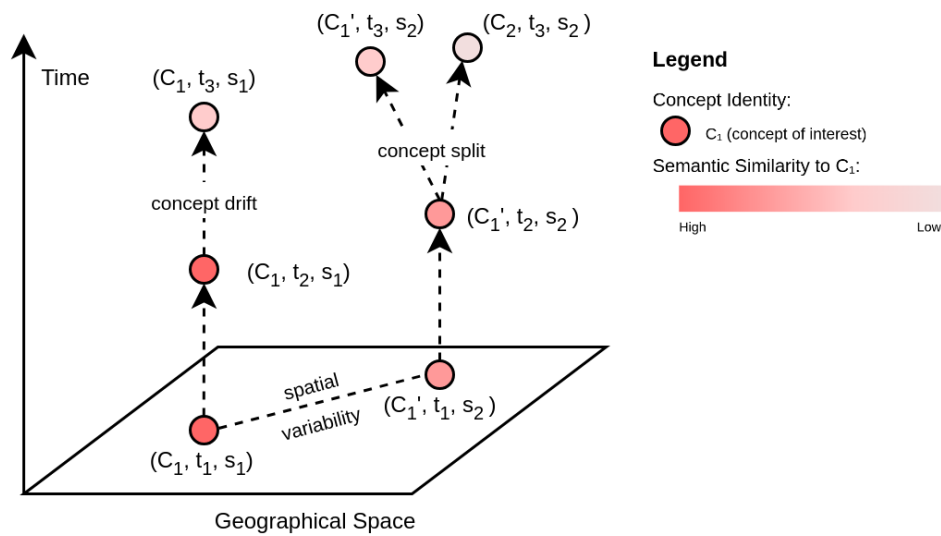


Figure 1 Representation of concepts drift and split over geographic space and time. The intensity of color indicates concept similarity.

3 Related Work

This section reviews existing work that provides means for measuring the latent semantics underlying words in their embeddings, with a spatio-temporal focus.

3.1 Spatial and Temporal Information in Word Embeddings

With the introduction of Word2Vec [29], word embeddings have revolutionized representation by converting words into dense vectors in a high-dimensional space,⁵ where semantically similar words are close to each other. Such representation enables a more flexible study of temporal and spatial variations in lexical semantics, offering an advantage over directly comparing different ontology versions. Early pre-trained word embeddings, such as GloVe [32], provide static representations, where each word is assigned a single vector, independent of context. Later, more advanced models like BERT [10] provide context-aware embeddings that capture more variations in meaning based on surrounding text using attention mechanisms.

Several studies [3, 20, 37] have explored the enrichment of word embeddings with temporal and spatial information. For example, Zhang et al. [54] focused on temporal counterpart search that detects semantically similar terms over time. The authors later also investigated the geographic variations in lexical semantics [55], e.g., showing that **typhoon** in Japan would be the most similar term to **hurricane** in the United States. Gong et al. [16] further extended this idea and proposed a model that conditions word embeddings on time *or* location (i.e., generating time- and location-specific embeddings). Their findings included word similarities over time (e.g., **bitcoin** in 2015 and **stocks** in 1992) and locations (e.g., **president** in the United States and **prime minister** in Canada). A few other studies in GIScience explored

⁵ Note that some literature uses the term “low-dimensional space” here when comparing the dimensionality to a one-hot encoding. We use the term here to signify that the resulting embeddings are in a, say, 300-dimensional vector space.

the representation learning of *places* via word embeddings, thereby using spatial information alone. Yan et al. [52] applied word-embedding techniques to learn embeddings of places based on their types and distances. Later, Zhai et al. [53] extended this approach to the representation learning of functional regions.

While these approaches captured variations in lexical semantics along one dimension effectively, they did not *jointly* consider spatial and temporal dimensions. As a result, they would fail to capture spatial-temporal lexical similarity, such as **chancellor** in Germany in 2010 and **prime minister** in the United Kingdom in 1980. Such similarity is centered in our study on spatio-temporal concept drift, which accounts for both dimensions at the same time.

3.2 Word Embedding Association Test

The Word Embedding Association Test (WEAT) [7], inspired by the Implicit Association Test in psychology, is a widely used method for quantifying semantic associations in word embeddings. It calculates association scores by comparing cosine similarities between two sets of target words (e.g., man and woman) and two sets of attribute words (e.g., doctor and gynecologist). For example, associations between \overrightarrow{man} and \overrightarrow{doctor} versus \overrightarrow{woman} and $\overrightarrow{gynecologist}$ can be assessed using vector arithmetic [6, 30], expressed as $\overrightarrow{man} - \overrightarrow{woman} \approx \overrightarrow{doctor} - \overrightarrow{gynecologist}$. The resulting WEAT score indicates the degree of association between the two groups in the embedding space. WEAT provides a standardized measure and allows for statistical significance testing of observed changes. By adapting this method, we can compute the cosine similarities between different temporal snapshots and geographical regions, e.g., (*hurricane* to *Mexico*, 2005) and (*typhoon* to *China*, 2015). Note that 2005 and 2015 are not treated as vectors themselves but rather indicate the time periods associated with these concept-region pairs. If in a vector space, $\overrightarrow{hurricane} - \overrightarrow{typhoon} \approx \overrightarrow{Mexico} - \overrightarrow{China}$, this would allow us to quantify concept changes across space and time and reveal geographic prototypes underlying word embeddings. However, applying WEAT to spatio-temporal analysis also presents challenges, particularly in maintaining statistical power when data is sparse across certain regions or time periods.

4 Case Study: Geographical Information Science Publications

We employ a scientometric dataset from Wu et al. [51] to detect the spatio-temporal concept drift in a real-world dataset. This dataset includes research publications in the field of GIScience from 1991 to 2020, sourced from Scopus⁶. As the dataset focuses on international journals and conferences that publish exclusively in English, all included publications are in English. Here, we explicitly focus on papers that mention locations in their abstracts, including geopolitical entities (GPEs) – such as countries, states, and cities – as well as nationalities (NORP), using the spaCy transformer-based named entity recognition pipeline⁷. Table 1 presents the summary statistics of the dataset after filtering for these papers. For reference, we provide the full names of conference and journal abbreviations in Appendix A.

⁶ <https://www.scopus.com/>

⁷ <https://spacy.io/models>

■ **Table 1** Summary statistics of research publications with location mentions in abstracts.

Type	Name	Time Range	Number of Papers	Number of Keywords
Conference	COSIT	1993-2019	22	100
Conference	GIScience	2006-2020	18	78
Journal	CEUS	1999-2020	622	3177
Journal	CaGIS	1991-2020	196	1002
Journal	EPB	1998-2020	352	1744
Journal	GeoI	1997-2020	61	305
Journal	IJGIS	2005-2020	550	2640
Journal	JGS	1996-2020	196	922
Journal	JOSIS	2010-2020	21	133
Journal	SCC	2003-2020	19	92
Journal	TGIS	2007-2020	50	237
Total		1991-2020	2107	10430

4.1 Spatio-Temporal Dimensions of Concepts

As with the *what*, *when*, and *where* questions, we argue that each concept has thematic, temporal, and spatial dimensions. In this dataset, we treat each keyword as signifying an individual concept⁸ and represent these three dimensions accordingly.

To represent the (1) **thematic dimension**, we use the associated abstract, which provides contextual information of each concept. Since Scopus is an abstract and citation database without guaranteed full-text access, abstracts – being more consistently available across publications – are a practical choice for large-scale analysis. For the (2) **temporal dimension**, we use the publication year of the paper associated with each concept. Lastly, for the (3) **spatial dimension**, we use OpenStreetMap Nominatim⁹ to geocode the identified locations and extract the corresponding country for sub-national locations (e.g., cities). For location mentions like “East African”, we retain them at the continent level. If multiple countries are mentioned in an abstract, we document all of them. After retrieving 2,112 publications with location mentions, we manually reviewed 16 unidentified cases, assigning them to the country level or removing them where necessary. This resulted in a final dataset of 2,107 publications.

Table 2 includes examples of abstracts with location mentions and the extracted countries (or regions). The distribution of the 10 most mentioned countries in publications within our dataset is visualized in the heatmap in Figure 2. From this heatmap, we can observe that the United States and China lead in the number of publications, followed by other English-speaking countries (e.g., the United Kingdom, Canada, and Australia) and several European countries. Along the temporal axis, we also observe a notable increase in publications since 2005 in this scientometric dataset.

4.2 Spatio-Temporal Concept Drift in Embedding Space

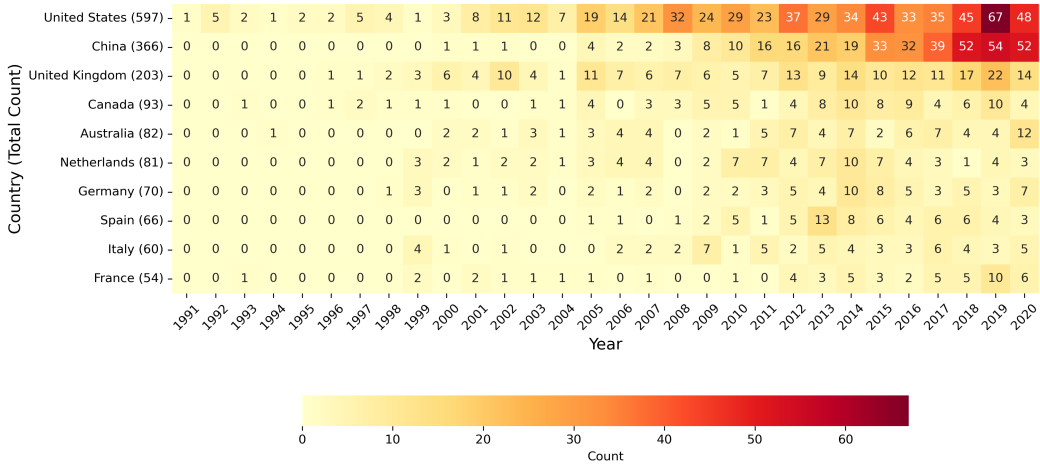
With the defined spatial and temporal dimensions of each concept, we leverage word embeddings to capture their variations in the thematic dimension. We employ the pre-trained SciBERT model [5], which is designed for scientific texts, to compute embeddings. We use

⁸ The distinction between a symbol and a concept can be explained using the triangle of reference [31].

⁹ <https://nominatim.org>

■ **Table 2** Exemplar location mentions and extracted countries in abstracts.

Year	Abstract Excerpt	Location Mentions	Country
1999	“We illustrate...based on the street pattern of a small French town.”	[French]	[France]
2008	“A dataset describing...in New York City is analyzed to...the technique.”	[New York City]	[United States]
2017	“Using Austria and Slovenia as a study area,...modified IL.”	[Austria, Slovenia]	[Austria, Slovenia]



■ **Figure 2** Distribution of publications for the 10 most mentioned countries over time.

context-aware representations of concepts in natural language, i.e., $(label(C), context(C))$ for a concept C , as discussed in Section 2.1. Here, we compute these two types of embeddings for concepts (in this case, keywords): (1) **label embedding**, which is static and derived from the keywords themselves, and (2) **context embedding**, which is context-aware and based on their associated abstracts with location mentions.

Additionally, we investigate the sensitivity of context embeddings to location mentions by computing (3) **context embedding without locations** as well. This embedding is derived from associated abstracts of a concept, where each identified location mention is replaced with the placeholder “[Location]”. For instance, in the first example in Table 2, the sentence would become “We illustrate...based on the street pattern of a small [Location] town.” This helps reveal whether explicit geospatial references influence the context-aware representation of concepts.

For each keyword/concept C in a given year t and country (or region) s , we first average its context embeddings across all relevant abstracts, to ensure a single embedding for each unique keyword-year-country combination. We then integrate this with the label embedding through a convex combination to obtain the composite embedding $C_{t,s}$, formulated as:

$$C_{t,s} = \alpha \cdot label(C) + (1 - \alpha) \cdot \frac{1}{|D_{t,s}|} \sum_{d \in D_{t,s}} context(C, d) \quad (1)$$

where $label(C)$ is the embedding of the keyword itself through its label; $context(C, d)$ is the embedding of the abstract in document d containing the keyword; $D_{t,s}$ is the set of documents that contain the keyword from year t and country s ; and $|D_{t,s}|$ is the number of such documents. The parameter α determines the weight assigned to the label versus context embeddings.

To quantify how a concept C drifts across different space-time combinations, we use cosine similarity between their respective composite embeddings. Given two concept representations, e.g., C_{t_1,s_1} and C_{t_2,s_2} , their similarity is computed as:

$$sim(C_{t_1,s_1}, C_{t_2,s_2}) = \frac{C_{t_1,s_1} \cdot C_{t_2,s_2}}{\|C_{t_1,s_1}\| \cdot \|C_{t_2,s_2}\|} \quad (2)$$

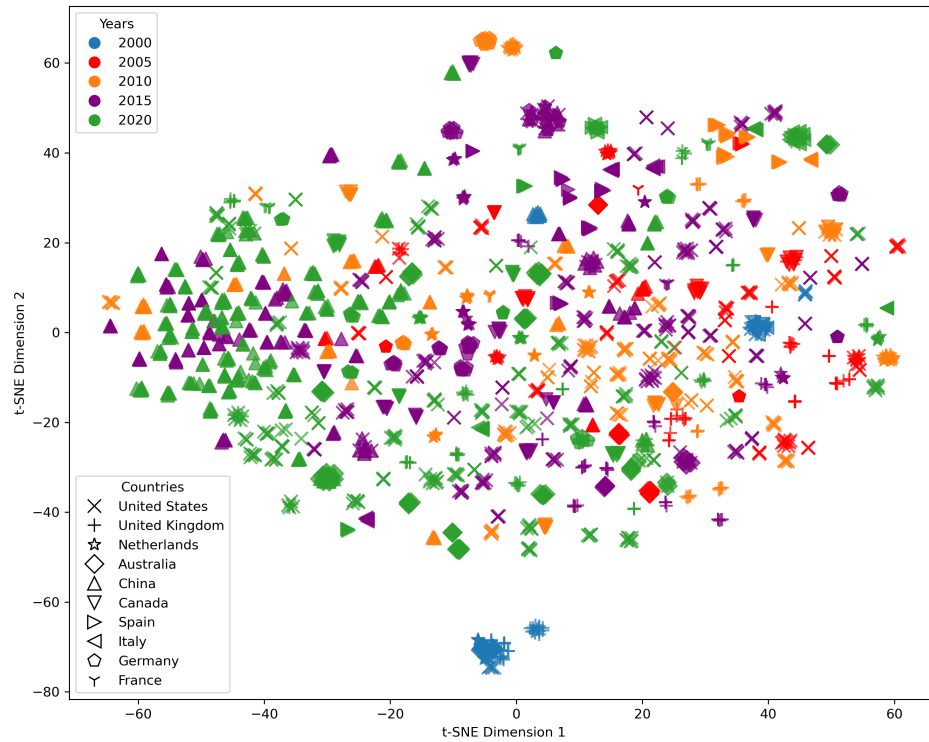
5 Results

To start, we visualize these keywords in the embedding space. Figure 3 shows the distribution of keyword embeddings across countries and selected years (2000, 2005, 2010, 2015, and 2020), generated using t-Distributed Stochastic Neighbor Embedding (t-SNE) [46]. The label embedding weight α is empirically set to 0.3 to place greater emphasis on the context embedding while retaining sufficient label information. Higher values of α tend to produce overly label-driven clusters, whereas lower values may cause semantically related keywords to diverge (see Appendix B for examples).

From the figure, we can observe a cluster of keywords with location mentions of China over the years (represented by triangles of different colors in the middle left of the figure), and those with location mentions of European countries like Germany and the Netherlands appear closer to each other.

The t-SNE visualization provides an overview of keyword distributions; we then look into how individual keywords move along their semantic trajectories across different countries over the years. Note that all keywords are standardized to lowercase and American spelling. They are also lemmatized and expanded to their full forms (e.g., *DEM* to *digital elevation model*), with the exception of *GIS*, which we retain as an abbreviation due to its ambiguous reference to *GI Science* or *GI System*. For each keyword, we quantify its spatio-temporal coverage by multiplying its time span (in years) by the number of unique countries it is associated with, yielding a coverage score to reflect both its temporal persistence and geographic distribution. Table 3 presents the top 10 keywords ranked by their spatio-temporal coverage. From these, we plot the semantic trajectories for selected keywords – *GIS*, *urban planning*, *spatial analysis*, and *cellular automaton* – in Figure 4, using principal component analysis (PCA) [1] to reduce the dimensionality of their embeddings.

From these trajectories, we can notice that the extracted embeddings of *GIS* (Figure 4a) are quite consistent across Italy, Germany, Australia, and the UK in the early years of 1999 and 2000. Afterward, English-speaking countries, including the UK, the US, Canada, and Australia, along with China, have their embeddings clustered together. In contrast, European countries, e.g., Spain, France, and Italy, form a separate cluster between 1999 and 2015. This reflects that these countries might take different approaches to GIS theories and applications. Contrary to *GIS*, the semantic trajectories for *urban planning* (Figure 4b) and *spatial analysis* (Figure 4c) vary significantly across different countries. This indicates that these two keywords exhibit strong region-specific embeddings, reflecting that the research under these two keywords in our scientometric dataset is potentially more influenced by local policies, socioeconomic conditions, and so on. Their variations across countries also suggest that, even based on the same theoretical foundation, the practical applications of



■ **Figure 3** A t-SNE visualization of keyword embeddings with selected years and countries.

these concepts can vary and lead to country-dependent interpretations. Compared with *urban planning* and *spatial analysis*, *cellular automaton* (Figure 4d) shows similar semantic trajectories and clustered embeddings across countries. This country-wise consistency is likely attributed to the stronger mathematical and computational foundations of *cellular automaton*, which makes it potentially less influenced by local policies or conditions. This observation also indicates a more widely shared understanding and development of theories and applications in cellular automaton.

■ **Table 3** The top 10 keywords by spatio-temporal coverage.

Keyword	Time Span	Unique Years	Unique Countries	Total Count	Coverage Score
GIS	1993-2020 (27)	26	45	138	1215
Geographic Information System	1994-2020 (26)	20	36	72	936
Remote Sensing	1995-2020 (25)	17	29	48	725
Land Use	1995-2020 (25)	20	26	55	650
Model	1999-2020 (21)	12	28	41	588
Urban Planning	1998-2020 (22)	15	25	47	550
Spatial Analysis	1998-2020 (22)	18	25	54	550
Cellular Automaton	2000-2020 (20)	19	25	63	500
Visualization	1997-2019 (22)	14	20	39	440
Cadastre	2001-2020 (19)	8	20	24	380

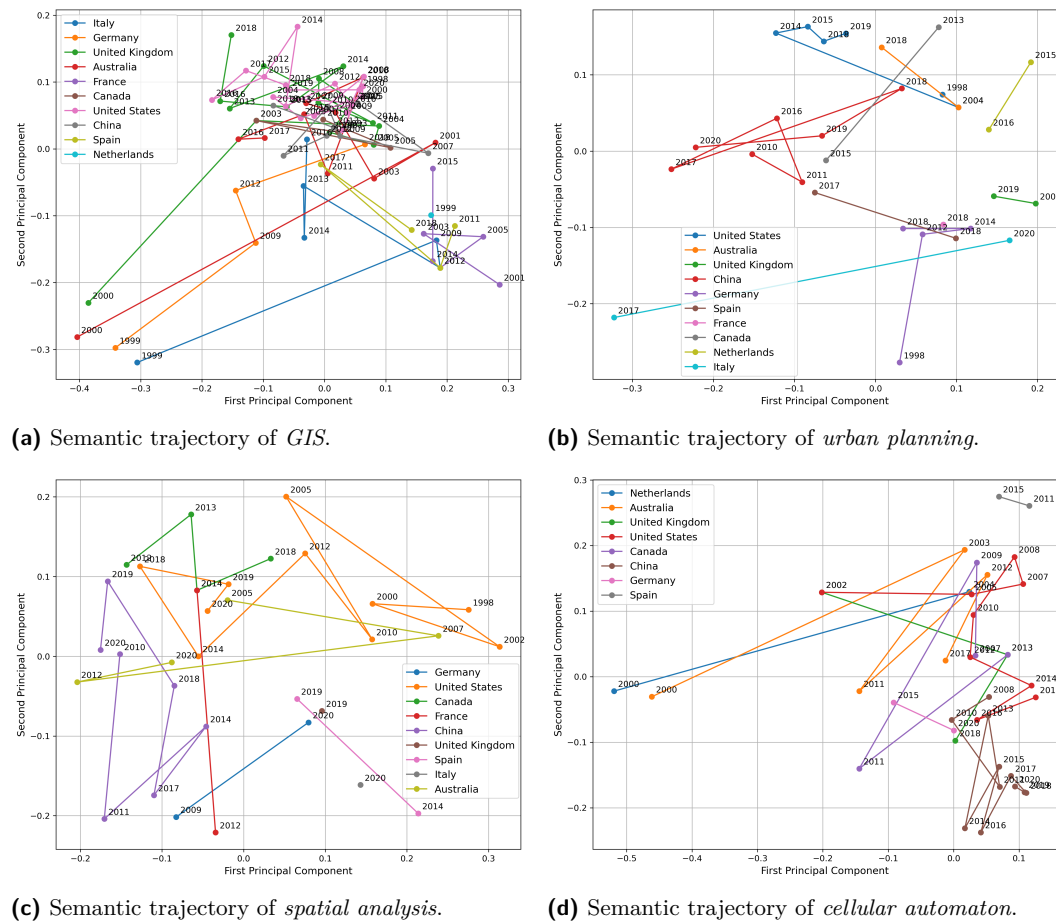


Figure 4 PCA visualization of the semantic trajectory of keywords by country over time. Only the top 10 countries are included for visual clarity. We use the first and the second principle components for PCA visualization.

While semantic trajectories trace how a single keyword (used as a proxy for an underlying concept) evolves over time and across countries, they do not capture how it relates to other keywords in semantic space. Table 4 presents selected examples of keywords in different country-year combinations and how their meanings evolve. For each unique keyword-country-year combination, we identify its three most similar keywords from the other years, calculated based on Equation 2. For *GIS* in Austria (1999), its closest semantic matches are found in the UK and Australia to itself and *integrated model* in the Netherlands. The temporal distance between these matches is small, suggesting that the conceptualization of GIS remained relatively stable across these countries during this period. In contrast, *machine learning* in France (2003) follows a different pattern. It shares the strongest similarities with *ontology* and *temporal management* in Spain (2012) but also aligns with *machine learning* in Czechia (2020). This indicates that the early machine learning concept was probably integrated into various domains of GIScience over time. Lastly, *urban planning* in China (2011 and 2019) show a rather location-stable pattern that the strongest similarities are all found within China across different years and with related planning keywords. This suggests a more internally consistent evolution of urban planning concepts within the Chinese research community. In contrast, *urban planning* in Australia (2004 and 2018) shows similarities across many

countries (the US, the UK, Netherlands, and Finland). This indicates a more dynamic and globally connected evolution of the underlying concept. We can infer from this comparison that the concept of *urban planning* may drift slower and more localized in China, while at a faster rate and more international in Australia.

■ **Table 4** Selected cases of the top three similar keywords and their similarity scores across countries and years.

Query Keyword	Top 3 Similar Keywords	Sim.
GIS (Austria, 1999)	GIS (United Kingdom, 2000)	0.931
	Integrated Model (Netherlands, 2000)	0.894
	GIS (Australia, 2000)	0.892
Machine Learning (France, 2003)	Ontology (Spain, 2012)	0.916
	Temporal Management (Spain, 2012)	0.916
	Machine Learning (Czechia, 2020)	0.909
Urban Planning (China, 2011)	Urban Planning (China, 2019)	0.942
	Geographic Information System (China, 2016)	0.939
	Planning Support System (China, 2020)	0.938
Urban Planning (China, 2019)	Urban Spatial Dynamic (China, 2020)	0.953
	Scenario Planning (China, 2020)	0.948
	Urban Land Use (China, 2020)	0.948
Urban Planning (Australia, 2004)	Urban Planning (Finland, 2020)	0.947
	Urban Planning (United States, 1998)	0.937
	Urban Planning (Netherlands, 2016)	0.936
Urban Planning (Australia, 2018)	Urban Data (United States, 2019)	0.928
	Urban Land Use Change (United Kingdom, 2014)	0.927
	Urban Scaling Law (Europe, 2020)	0.924

Finally, we perform a sensitivity analysis on countries with at least 100 associated keywords to evaluate the impact of explicit location mentions on the extracted context embeddings. Using a two-tailed permutation test (with 1,000 permutations), we find that the average cosine distance ($1 - \text{cosine similarity}$) between embeddings with and without location mentions (0.010) is significantly smaller than what would be expected by chance (permutation mean = 0.242, std = 0.0005, $p < 0.001$). This indicates that explicit location mentions have minimal impact on the semantic representation of concepts in our case study. This is likely because geographical context is implicitly encoded in the text. We discuss this in more detail in the discussion section.

6 Discussion

This section discusses the challenges of defining spatio-temporal dimensions of concepts and the potential biases introduced in our case study. We also discuss findings from the sensitivity analysis, outline future research directions and the implications of spatio-temporal concept drift for ontology learning with large language models (LLMs).

6.1 Challenges in Defining Spatio-Temporal Dimensions of Concepts

Understanding spatio-temporal concept drift in scientific texts requires linking keywords (and their underlying concepts) to geographic locations and time, but this process inherently introduces biases.

In our case study, we use publication year as the temporal dimension of a concept. However, the publication year could be different from the actual study period (e.g., a paper published in 2010 on East Africa in the 1970s). We attribute spatial dimensions of concepts (keywords) based on location mentions in their associated abstracts. However, not all location mentions correspond to the actual study area; some may appear as examples, comparisons, or even counterexamples. We then aggregate these locations to use the country as the spatial unit of concept drift, overlooking regional variations within the country level. Take the concept of *urban planning* as an example. Its interpretation could differ significantly for New York City (e.g., a walkable, transit-oriented city) versus Los Angeles (e.g., a car-centric city) over the years. These regional disparities would become particularly pronounced in larger countries (e.g., the United States) with diverse geographic and socioeconomic conditions. Our spatial aggregation approach implicitly assumes concept homogeneity at the country level, which introduces biases into the learned embeddings.

Future work could explore improved spatio-temporal scoping techniques to capture study periods and areas more accurately; it should also include different spatial levels – cities, countries, and continents – to measure variations.

6.2 Sensitivity to Location Removal in Context Embeddings

Our sensitivity analysis reveals varying effects across countries when removing location mentions. For example, keywords associated with Japan show slightly larger differences in their embeddings, though the overall differences remained small. Here, several factors may complicate the interpretation of these results. First, the dataset has a substantial imbalance, with publications mentioning the United States far outnumbering those that mention other countries. When extracting unique country-year combinations, this imbalance leads to sparse samples for less represented countries, thus masking meaningful patterns. Second, we did not account for the proportion of location mentions within abstracts, e.g., some abstracts contain a list of study areas, while others mention one location briefly. The observed differences in embeddings with and without locations may be due to the removal of more contextual information rather than an inherent sensitivity to geographic reference. These factors need a more fine-grained analysis to quantify the impact of explicit location mentions on embedding representations in future work.

6.3 Implication for Ontology Learning with Large Language Models

Since large language models (LLMs) become more commonly used for ontology learning tasks [27, 2, 41], we need to ensure geographic and temporal variations in knowledge are accounted for to have more context-aware representations. Current LLMs are trained on vast corpora of text that usually lack explicit spatial and temporal structuring [13], which would likely overlook the spatio-temporal variations in concept representations unless specified in the prompt. Our findings show that concepts in scientific texts evolve differently across geographic space and over time. This suggests that ontologies derived from LLMs may inherit hidden geographic biases. For instance, when an LLM processes the concept of *smart city*, its interpretation might be overly influenced (and represented) by temporally and regionally dominant implementations (e.g., the recent decade in Singapore) if the training data is dominated by publications from this region and time period.

Our observations show that concepts in scientific texts can vary across geographic space and time, and suggest the need for a more context-aware mechanism when using LLMs for ontology learning. This could be achieved with region-specific knowledge validation

and/or the development of geographically aware prompting strategies. To capture the spatio-temporal dynamics in scientific concepts, future work could include the design of more few-shot learning approaches, where examples are carefully selected to represent diverse temporal and geographic interpretations of concepts.

7 Conclusions

Space and time are central to the study of geography and GIScience. These dimensions not only shape our daily physical interactions of *when* and *where*, but also influence the abstract representation of concepts in scientific knowledge. With the ever-increasing volume of research publications, we need methods to better structure concepts embedded in scientific research for organization and retrieval purposes. Encoded in an ontology, we could also account for the spatio-temporal dynamics of concepts, which are constantly evolving – often at varying rates across regions – due to technology and societal changes, for more effective research data management (RDM).

In this work, we introduce the notion of spatio-temporal concept drift. We complement previous work on concept drift by including the spatial dimension, and propose a novel approach using word embedding techniques to capture this drift over space and time. Using a scientometric dataset in the field of GIScience, we demonstrate that keywords (used as proxies for underlying concepts) show varying drift patterns over time and across countries. Spatially grounded concepts, such as *urban planning* (as compared to *cellular automaton*), can have substantial differences in meanings for different countries and over time.

The implications of this work extend beyond improving the understanding of concepts in scientific texts to enhancing FAIR-based RDM systems. This reminds us, for example, that concepts like *cellular automaton* may require less user intervention, while concepts like *urban planning* may need query enrichment to account for local and temporal variation to better match a user’s keyword. Given the observed spatio-temporal concept drift and the increasing use of ontology learning with large language models (LLMs), we also suggest that LLM-based ontology learning mechanisms should explicitly account for the spatial and temporal dimensions of concept representation. Making RDM systems and ontology learning approaches more sensitive to these variations will help improve retrieval and maintain the relevance of knowledge in scientific texts.

References

- 1 Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010. doi:10.1002/wics.101.
- 2 Hamed Babaei Giglou, Jennifer D’Souza, and Sören Auer. LLMs4OL: Large language models for ontology learning. In *International Semantic Web Conference*, pages 408–427. Springer, 2023. doi:10.1007/978-3-031-47240-4_22.
- 3 David Bamman, Chris Dyer, and Noah A Smith. Distributed representations of geographically situated language. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 828–834, 2014. doi:10.3115/v1/p14-2134.
- 4 Lawrence Barsalou. Concepts and meaning. In L. Barsalou, W. Yeh, B. Luka, K. Olseth, K. Mix, and L. Wu, editors, *Chicago Linguistic Society 29: Papers From the Parasession on Conceptual Representations*, pages 23–61. University of Chicago, 1993.
- 5 Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In *Conference on Empirical Methods in Natural Language Processing*, 2019. doi:10.18653/v1/D19-1371.

- 6 Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016. doi:10.48550/arXiv.1607.06520.
- 7 Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017. doi:10.1126/science.aal4230.
- 8 Giuseppe Capobianco, Danilo Cavaliere, Sabrina Senatore, et al. Ontodrift: a semantic drift gauge for ontology evolution monitoring. In *CEUR Workshop Proceedings*, volume 2821, pages 1–10. CEUR-WS, 2020. URL: <https://ceur-ws.org/Vol-2821/paper1.pdf>.
- 9 Christophe Claramunt. Ontologies for geospatial information: Progress and challenges ahead. *Journal of Spatial Information Science*, 20:35–41, 2020. doi:10.5311/JOSIS.2020.20.666.
- 10 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019. doi:10.18653/v1/N19-1423.
- 11 Stephanie Duce and Krzysztof Janowicz. Microtheories for spatial data infrastructures-accounting for diversity of local conceptualizations at a global level. In *Geographic Information Science: 6th International Conference, GIScience 2010, Zurich, Switzerland, September 14-17, 2010. Proceedings 6*, pages 27–41. Springer, 2010. doi:10.1007/978-3-642-15300-6_3.
- 12 Max J Egenhofer and David M Mark. Naive geography. In *Spatial Information Theory A Theoretical Basis for GIS: International Conference COSIT'95 Semmering, Austria, September 21-23, 1995 Proceedings 2*, pages 1–15. Springer, 1995. doi:10.1007/3-540-60392-1_1.
- 13 Fahim Faisal and Antonios Anastasopoulos. Geographic and geopolitical biases of language models. In Duygu Ataman, editor, *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 139–163, Singapore, December 2023. Association for Computational Linguistics. doi:10.18653/v1/2023.mrl-1.12.
- 14 Antske Fokkens, Serge Ter Braake, Isa Maks, Davide Ceolin, et al. On the semantics of concept drift: Towards formal definitions of semantic change. *Drift-a-LOD@ EKAW*, 2016. URL: https://ceur-ws.org/Vol-1799/Drift-a-LOD2016_paper_2.pdf.
- 15 Aldo Gangemi, Nicola Guarino, Claudio Masolo, Alessandro Oltramari, and Luc Schneider. Sweetening ontologies with DOLCE. In *International conference on knowledge engineering and knowledge management*, pages 166–181. Springer, 2002. doi:10.1007/3-540-45810-7_18.
- 16 Hongyu Gong, S. Bhat, and Pramod Viswanath. Enriching word embeddings with temporal and spatial information. In *Conference on Computational Natural Language Learning*, 2020. doi:10.18653/v1/2020.conll-1.1.
- 17 Nicola Guarino. Formal ontology, conceptual analysis and knowledge representation. *International journal of human-computer studies*, 43(5-6):625–640, 1995. doi:10.1006/ijhc.1995.1066.
- 18 Ramanathan Guha, Rob McCool, and Eric Miller. Semantic search. In *Proceedings of the 12th international conference on World Wide Web*, pages 700–709, 2003. doi:10.1145/775152.775250.
- 19 Prashant Gupta and Mark Gahegan. Categories are in flux, but their computational representations are fixed: That’s a problem. *Transactions in GIS*, 24(2):291–314, 2020. doi:10.1111/tgis.12602.
- 20 William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, 2016. doi:10.18653/v1/P16-1141.
- 21 Krzysztof Janowicz, Song Gao, Grant McKenzie, Yingjie Hu, and Budhendra Bhaduri. GeoAI: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond, 2020. doi:10.1080/13658816.2019.1684500.

- 22 Krzysztof Janowicz, Martin Raubal, and Werner Kuhn. The semantics of similarity in geographic information retrieval. *Journal of Spatial Information Science*, 2:29–57, 2011. doi:10.5311/JOSIS.2011.2.3.
- 23 Werner Kuhn. Semantic reference systems. *International Journal of Geographical Information Science*, 17(5):405–409, 2003. doi:10.1080/1365881031000114116.
- 24 Werner Kuhn, Martin Raubal, and Peter Gärdenfors. Cognitive semantics and spatio-temporal ontologies. *Spatial Cognition & Computation*, 7(1):3–12, 2007. doi:10.1080/13875860701337835.
- 25 Stephen C Levinson. Language and space. *Annual review of Anthropology*, 25(1):353–382, 1996. doi:10.1146/annurev.anthro.25.1.353.
- 26 Maurice Lineman, Yuno Do, Ji Yoon Kim, and Gea-Jae Joo. Talking about climate change and global warming. *PloS one*, 10(9):e0138996, 2015. doi:10.1371/journal.pone.0138996.
- 27 Huu Tan Mai, Cuong Xuan Chu, and Heiko Paulheim. Do LLMs really adapt to domains? an ontology learning perspective. In *International Semantic Web Conference*, pages 126–143. Springer, 2024. doi:10.1007/978-3-031-77844-5_7.
- 28 David M. Mark, Barry Smith, and Barbara Tversky. Ontology and geographic objects: An empirical study of cognitive categorization. In *Conference On Spatial Information Theory*, pages 283–298, 1999. doi:10.1007/3-540-48384-5_19.
- 29 Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*, 2013. doi:10.48550/arXiv.1301.3781.
- 30 Malvina Nissim, Rik van Noord, and Rob van der Goot. Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, 46(2):487–497, 2020. doi:10.1162/coli_a_00379.
- 31 Charles Kay Ogden and Ivor Armstrong Richards. *The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism*. Harcourt, Brace & World, Inc., 1923. doi:10.2307/2015195.
- 32 Jeffrey Pennington, Richard Socher, and Christopher D Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. doi:10.3115/v1/d14-1162.
- 33 Dharmen Punjani, Kuldeep Singh, Andreas Both, Manolis Koubarakis, Iosif Angelidis, Konstantina Bereta, Themis Beris, Dimitris Bilidas, Theofilos Ioannidis, Nikolaos Karalis, et al. Template-based question answering over linked geospatial data. In *Proceedings of the 12th workshop on geographic information retrieval*, pages 1–10, 2018. doi:10.1145/3281354.3281362.
- 34 Martin Raubal. Formalizing conceptual spaces. In *Formal ontology in information systems, proceedings of the third international conference (FOIS 2004)*, volume 114, pages 153–164. Citeseer, 2004.
- 35 Martin Raubal. Representing concepts in time. In *Spatial Cognition VI. Learning, Reasoning, and Talking about Space: International Conference Spatial Cognition 2008, Freiburg, Germany, September 15-19, 2008. Proceedings 6*, pages 328–343. Springer, 2008. doi:10.1007/978-3-540-87601-4_24.
- 36 M Andrea Rodriguez and Max J. Egenhofer. Determining semantic similarity among entity classes from different ontologies. *IEEE transactions on knowledge and data engineering*, 15(2):442–456, 2003. doi:10.1109/TKDE.2003.1185844.
- 37 Maja Rudolph and David Blei. Dynamic embeddings for language evolution. In *Proceedings of the 2018 world wide web conference*, pages 1003–1011, 2018. doi:10.1145/3178876.3185999.
- 38 Christoph Schlieder. Digital heritage: Semantic challenges of long-term preservation. *Semantic Web*, 1(1-2):143–147, 2010. doi:10.3233/SW-2010-0013.
- 39 Angela Schwering. Approaches to semantic similarity measurement for geo-spatial data: a survey. *Transactions in GIS*, 12(1):5–29, 2008. doi:10.1111/j.1467-9671.2008.01084.x.

- 40 Meilin Shi, Krzysztof Janowicz, Zilong Liu, Mina Karimi, Ivan Majic, and Alexandra Fortacz. Defining concept drift and its variants in research data management: A scientometric case study on geographic information science. *Transactions in GIS*, 29(3):e70058, 2025. doi:10.1111/tgis.70058.
- 41 Cogan Shimizu and Pascal Hitzler. Accelerating knowledge graph and ontology engineering with large language models. *Journal of Web Semantics*, page 100862, 2025. doi:10.1016/j.websem.2025.100862.
- 42 Barry Smith and David M Mark. Geographical categories: an ontological investigation. *International journal of geographical information science*, 15(7):591–612, 2001. doi:10.1080/13658810110061199.
- 43 Barry Smith and David M Mark. Do mountains exist? towards an ontology of landforms. *Environment and Planning B: Planning and Design*, 30(3):411–427, 2003. doi:10.1068/b12821.
- 44 Thanos G Stavropoulos, Stelios Andreadis, Efstratios Kontopoulos, and Ioannis Kompatsiaris. SemaDrift: A hybrid method and visual tools to measure semantic drift in ontologies. *Journal of Web Semantics*, 54:87–106, 2019. doi:10.1016/j.websem.2018.05.001.
- 45 Wolfgang G Stock. Concepts and semantic relations in information science. *Journal of the American Society for Information Science and Technology*, 61(10):1951–1969, 2010. doi:10.1002/asi.21382.
- 46 Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.
- 47 Saskia Van Putten, Carolyn O’Meara, Flurina Wartmann, Joanne Yager, Julia Villette, Claudia Mazzuca, Claudia Bieling, Niclas Burenhult, Ross Purves, and Asifa Majid. Conceptualisations of landscape differ across european languages. *Plos one*, 15(10):e0239858, 2020. doi:10.1371/journal.pone.0239858.
- 48 Stella Verkijk, Ritten Roothaert, Romana Pernisch, and Stefan Schlobach. Do you catch my drift? on the usage of embedding methods to measure concept shift in knowledge graphs. In *Proceedings of the 12th Knowledge Capture Conference 2023*, pages 70–74, 2023. doi:10.1145/3587259.3627555.
- 49 Shenghui Wang, Stefan Schlobach, and Michel Klein. Concept drift and how to identify it. *Journal of Web Semantics*, 9(3):247–265, 2011. doi:10.1016/j.websem.2011.05.003.
- 50 Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The FAIR guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016. doi:10.1038/sdata.2016.18.
- 51 Xiaohuan Wu, Weihua Dong, Lun Wu, and Yu Liu. Data and Code for "Research Themes of Geographical Information Science during 1991–2020: A Retrospective Bibliometric Analysis", 2022. doi:10.6084/m9.figshare.19242654.v1.
- 52 Bo Yan, Krzysztof Janowicz, Gengchen Mai, and Song Gao. From ITDL to Place2Vec: Reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts. In *Proceedings of the 25th ACM SIGSPATIAL international conference on advances in geographic information systems*, pages 1–10, 2017. doi:10.1145/3139958.3140054.
- 53 Wei Zhai, Xueyin Bai, Yu Shi, Yu Han, Zhong-Ren Peng, and Chaolin Gu. Beyond Word2vec: An approach for urban functional region extraction and identification by combining Place2vec and POIs. *Computers, Environment and Urban Systems*, 74:1–12, 2019. doi:10.1016/j.compenvurbsys.2018.11.008.
- 54 Yating Zhang, Adam Jatowt, Sourav S Bhowmick, and Katsumi Tanaka. The past is not a foreign country: Detecting semantically similar terms across time. *IEEE Transactions on Knowledge and Data Engineering*, 28(10):2793–2807, 2016. doi:10.1109/TKDE.2016.2591008.
- 55 Yating Zhang, Adam Jatowt, and Katsumi Tanaka. Is tofu the cheese of asia?: Searching for corresponding objects across geographical areas. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1033–1042, 2017. doi:10.1145/3041021.3055132.

A Conference and Journal Abbreviation Reference

■ **Table 5** Full names and abbreviations of selected conferences and journals.

Conference/Journal Name	Abbreviation
International Conference on Spatial Information Theory	COSIT
International Conference on Geographic Information Science	GIScience
Computers, Environment and Urban Systems	CEUS
Cartography and Geographic Information Science	CaGIS
Environment and Planning B: Urban Analytics and City Science	EPB
GeoInformatica	GeoI
International Journal of Geographical Information Science	IJGIS
Journal of Geographical Systems	JGS
Journal of Spatial Information Science	JOSIS
Spatial Cognition & Computation	SCC
Transactions in GIS	TGIS

B Sensitivity Analysis of the Label Embedding Weight

■ **Table 6** Top similar keywords retrieved under different values of the label embedding weight α , along with their similarity scores. The value of $\alpha = 0.3$ is used in the case study in this paper. Note that the examples are included post hoc to illustrate the qualitative effects of different α values.

Weight	Query Keyword	
	Urban Planning (Australia, 2018)	Climate Change (US, 2020)
$\alpha = 0.1$	urban scaling law (Europe, 2020): 0.925	climate change (US, 2014): 0.913
	zipf's law for city (Europe, 2020): 0.924	sea level rise (US, 2019): 0.910
	land use (Europe, 2020): 0.923	storm surge inundation (US, 2019): 0.909
	population density (Europe, 2020): 0.922	lidar (US, 2008): 0.909
	radial analysis (Europe, 2020): 0.922	greening scenario (US, 2018): 0.909
$\alpha = 0.3$	urban data (US, 2019): 0.928	climate change (US, 2014): 0.938
	urban land use change (UK, 2014): 0.927	climate change (UK, 2018): 0.919
	urban scaling law (Europe, 2020): 0.924	sea level rise (US, 2019): 0.912
	urban planning (US, 2019): 0.924	urban heat island (US, 2018): 0.911
	residential mobility (UK, 2014): 0.921	seasonal impact (US, 2018): 0.911
$\alpha = 0.5$	urban planning (US, 2019): 0.959	climate change (US, 2014): 0.967
	urban planning (Brazil, 2003): 0.948	climate change (UK, 2018): 0.956
	urban planning (Poland, 2017): 0.948	climate change (UK, 2012): 0.947
	urban planning (Spain, 2017): 0.947	climate change (US, 2013): 0.947
	urban planning (Netherlands, 2016): 0.943	climate change (US, 2015): 0.946

The Inherent Structure of Experiments as a Constraint to Spatial Analysis and Modeling

Simon Scheider   

Department of Human Geography and Spatial Planning, Utrecht University, The Netherlands

Judith A. Verstegen   

Department of Human Geography and Spatial Planning, Utrecht University, The Netherlands

Abstract

We argue that in order to justify a modeling approach for a particular purpose, we need to better understand the experimental structure that is supposed to be represented by a given model application. For this purpose, we introduce a logic for specifying causal as well as spatio-temporal experiments, based on which we reinterpret Sinton's structure of spatial information from a pragmatic, experimental viewpoint. We illustrate the use of this logic based on a landuse modeling example, showing to what extent remote sensing and simulation approaches can be justified by decomposing the example into experiments required for answering its main question.

2012 ACM Subject Classification General and reference → Experimentation; Information systems → Spatial-temporal systems

Keywords and phrases pragmatic Logic, experimental Norms, spatio-temporal Models

Digital Object Identifier 10.4230/LIPIcs.GIScience.2025.17

Funding *Simon Scheider*: This research was funded by the European Research Council (ERC) under the Horizon Europe programme (Grant Agreement No. 101170816).

1 Introduction

Experiments are fundamental to science. They not only serve to generate empirical knowledge, but also constrain how information sources are used in analysis and modeling to ensure valid results. They provide a basis for justification of knowledge and trust in scientific insights. Understanding experimental practice thus illuminates scientific methodology bottom-up, i.e., from study design and data acquisition to the construction of theoretical and computational models for addressing scientific questions [28, 23].

While machine learning based GeoAI modeling techniques [14] can simplify the design of complex models, our understanding of the experimental basis of the knowledge that is produced with such models still remains limited, in particular when deciding whether a given model can support a given claim or not [22]. Consider the example of land use change in Brazil, where increased demand for agricultural commodities such as bioethanol may drive deforestation. The process is complex: increased demand stimulates sugarcane expansion, yet sugarcane rarely replaces forests directly [1]. Instead, it displaces pastures, which then encroach upon forests (Fig. 1). Additional indirect effects arise from competing land uses, such as sugar and beef production.

Some studies claim to be able to detect and predict such indirect land use change via remote sensing [2], while others challenge this claim [27]. While remote sensing is a powerful tool for finding the visible traces of land use change, the images cannot directly reveal the causal mechanisms behind them. Assessing the effects of increased bioethanol demand, including indirect effects, requires a causal model that simulates controlled intervention experiments. Only in a model where certain invisible factors such as demand can be artificially controlled, fixed or left free for such a large system, we can compare two (with and



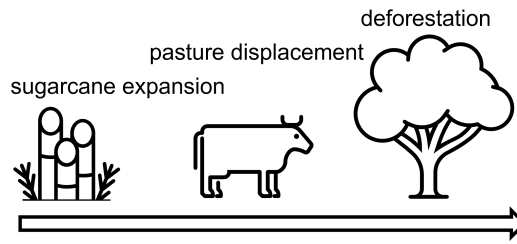
© Simon Scheider and Judith A. Verstegen;
licensed under Creative Commons License CC-BY 4.0

13th International Conference on Geographic Information Science (GIScience 2025).

Editors: Katarzyna Sila-Nowicka, Antoni Moore, David O'Sullivan, Benjamin Adams, and Mark Gahegan;
Article No. 17; pp. 17:1–17:17



Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



■ **Figure 1** How sugarcane expansion may cause deforestation.

without intervention) or more *possible* progressions of a process to find out the effects of an intervention. In contrast to remote sensing images, spatial simulation models, such as raster based land use change models, enable such reasoning [27]. Why is that? The disagreement in the community seems not merely related to model selection but to a deeper confusion about the types of experiments that different models can meaningfully represent.

Our scientific goal is thus fundamental: to clarify the role of experiments in the context of spatio-temporal modeling. This involves, on the one hand, understanding the *structure* of experiments – that is, what needs to be fixed, controlled, and measured – and how they can be *performed*. On the other hand, it requires understanding how we can *interpret* modeling purposes – namely, the questions a model is supposed to answer – in terms of such experiments. We argue that this kind of knowledge – *pragmatic knowledge*¹ [22] – is essential for interpreting models. Since models constrain the kinds of experiments they can represent, it is our pragmatic knowledge of the underlying experiment that allows us to judge whether a given spatio-temporal model is valid for a particular purpose. In recent work [23], we have suggested a way of understanding modeling purposes in terms of questions that reflect such spatio-temporal experiments, following insights on how the inherent structure of spatial information is a constraint to analysis, as suggested by David Sinton in 1978 [24]. However, while Sinton’s original idea of “attributes” “held constant”, “being controlled” or “measured” has inspired GIScientists to suggest corresponding geodata- and conceptual models [7, 3, 15], it remains underdeveloped from a theoretic point of view [5]. The idea has neither been rethought from the perspective of experimental design and causality, nor from a viewpoint of pragmatics². From this standpoint, we address the following key questions:

- **Q.** *What is the role of experiments in spatio-temporal modeling?*
- **Q A.** *What constitutes a spatio-temporal experiment?*
- **Q B.** *How is knowledge about the structure of experiments inherent in spatio-temporal modeling?*
- **Q C.** *Which types of spatio-temporal experiments need to be distinguished when answering questions with a model?*

For this purpose, we develop a pragmatic approach to experimental knowledge, drawing on the methodical constructivist school of philosophy [17, 11, 18]. According to these scholars [16, 10, 13], an *experiment* is an *action* that *implements* a situation (*condition*),

¹ The notion of pragmatics originates in linguistics, particularly in speech act theory. However, pragmatic methodology has far broader implications, placing action at the center of knowledge production [12].

² Our title is therefore rephrasing Sinton’s paper emphasizing the role of experiments.

initiates a *process* (the latter not being an action), and *observes* the resulting situation (*measure*). We suggest a logic of experimental knowledge to make explicit the structure of experiments underlying spatio-temporal models. To this end, we introduce a formal grammar of situations in Sec. 2, which serves to construct the knowledge claims that must be supported by experiments (Sec. 3). Our pragmatic logic is based on the work of the logician Paul Lorenzen and aligns with modern causal theory [20, 28]. We then place Sinton's ideas on firmer pragmatic grounds by introducing classes of spatio-temporal experiments in Sec. 4. Finally, in Sec. 5, we demonstrate how our theory can be used to decompose the land use change example above in terms of its inherent experiments. Based on this we justify a simulation modeling and reject a purely remote sensing-based approach.

2 A pragmatic grammar for situations and goals

In this section, we introduce a grammar for a pragmatic language following Lorenzen [17, 18] about situations underlying experiments, including actions, processes and states, as well as goals and imperatives which can be used to formulate requests. The language is explained with example sentences, and specified in terms of a basic EBNF syntax:

*rule*name : *expression*

where *expression* may consist of words for literals (“hello world”) or terms (without quotes) substitutable by further expressions. Expressions can be sequences (A B), alternatives (A | B) or repetitions (A?) (zero or one) of such words. A string is parsed by applying rules recursively to words in a sequence.

2.1 Predicators and nominators for things

We use words for kinds of things (*predicators*) and individual things of some kind (*nominators*). In addition to predicators for space and time which range over individual locations and moments in time (in spatial and temporal reference systems), we use the possibility of forming *amounts of space and time* [25], such as regions and time intervals. The former can be used to talk about the amount of space occupied by certain things. Similar predicators we use for amounts of stuff or objects [23]. Furthermore, we call all these predicators for space, objects, stuff, and their amounts *endurances*, meaning that they play a particular role in describing situations: they can *change in time*, whereas occurrences are the things that are *going on* in time, reflecting a common distinction in information ontology.

► Grammatical rule 1.

object : “house” | “river” | “ball” | ... | *person*

stuff : “energy” | *matter* | “heat” | ...

matter : “water” | “gold” | ...

portion : “amount of” (*object* | *stuff* | *space*)

endurance : *object* | *stuff* | *portion* | *space*

thing : *time* | *endurance*

predicator : *thing* | *occurrence*

Nominators allow us to refer to *particular things*, either by introducing names, or by using (in a common situation of speech) indicators (“this”, “that”) together with predicators:

► Grammatical rule 2.

here, there : “this” *space*

now, then : “this” *time*

home : “this” *house*

In the following we use various nominators for each predictor above, including names for persons, objects etc.

2.2 Occurrences, actions, situations and claims

Other predictors stand for different *occurrences*, to say what “goes on” with things. We distinguish dynamic from static occurrences using *process predictors* (involving some change of a situation that happens at a moment in time) and *state predictors* (involving some situation is static at a moment in time). Furthermore, we use a special class of predictors for talking about what can be done (*do-predictor*):

► **Grammatical rule 3.** *occurrence* : *process* / *state* / *do-predictor*

► **Grammatical rule 4.** *process* : “generate” / “stumble” / “rain” / “grow” / ...

► **Grammatical rule 5.** *state* : “stay” / “linger” / “rest” / ...

Do-predictors are distinct from other occurrences, since they stand for kinds of actions that can be attributed to the persons performing them, including their purposes [9, 12]:

► **Grammatical rule 6.** *do-predictor* : “make” / “measure” / “run” / “stay” / “drink” / “use” / ...

The copula κ is used to form situations with occurrences, to say that some occurrence has happened, and π to form situations with do-predictors, to denote action performances:

► **Grammatical rule 7.**

κ : “is” / “are”

π : “do” (“es”)?

happening : (at)? (time-nominator)? κ *occurrence* (“ing”? (appredictor)?

performance : (at)? (time-nominator)? π

action : *performance do-predictor* (“ing”? (appredictor)?

Appredictors are expressions that further specify the occurrence, which may use prepositions together with nominators. A *happening* uses a temporal nominator and the copula κ with some predictor for occurrences. For example:

“at that time is raining this amount of water”

“now is growing”

In a similar way, we use the copula π for reporting on *action performances*:

“at that time does stay at this house”

“now does run home”

Note that we can always interpret an action performance as if it was a process, i.e., a behavior [28], since do-predictors are occurrences. *Situations* are either happenings or actions that are controlled by endurance nominators, referring to those things to which this happens/who control the action. In particular, we require a person in control of actions:

► **Grammatical rule 8.**

situation : *endurance-nominator happening* / (person-nominator)? *action*

For example:

“this person at this time does stay at this house”

“here at this time is raining this amount of water”

“this tree now is growing”

The distinctive role of situations, which are sorts of *time-dependent* propositions, has been recognized early on in artificial intelligence, where they are called *fluents* [19].

► **Grammatical rule 9.** *proposition* : *situation* / ...

Propositions are used to make defensible *claims*. From a pragmatic perspective, the latter are *speech acts*, actions that can be performed by persons in a dialogue. To be able to express such acts, we introduce a way of saying that someone makes a claim using any proposition formed from the grammar above.

► **Grammatical rule 10.** *claim* : (*person-nominator*)? *performance* “(” *proposition* “)”

For example, Nora now makes the claim that it will be raining tomorrow:

“Nora now π (here tomorrow is raining)”

2.3 Goals and imperatives

Goals are propositions intended by persons. They can be wished without ever pursuing an action (wishful thinking), but in the more practically relevant cases, we talk about goals that actually can be pursued via actions. We form goals from propositions using a conjunction “such that” or \models . For example, if I am traveling, I might wish to be at home at a certain time:

“such that I then do stay at home”

We can distinguish goals based on what kind of proposition is used. Whenever we are using a situation as a goal, we are wishing that the latter may come about:

► **Grammatical rule 11** (goals).

\models : “*such that*”

goal : \models *situation* / ...

An example for a *modificative* goal is my wish to be at home (above), meaning a modification of the place at which I am staying. Imperatives are speech acts that prompt some action from a person. This can be expressed either by indicating the action directly, or by requesting a goal and leaving the action that *implements* the goal open to the person addressed. In order to express imperatives, we use the copula !:

► **Grammatical rule 12.** *imperative* : (*person-nominator*)? “!” (*action* / *goal*)

For example, a mother may request from her daughter Nora to be at home in time for dinner:

“Nora ! \models at this time are having dinner”

“Peter ! at this time do cycle home”

The first imperative is a request to *bring about some situation* using some modificative goal³. This leaves it open to Nora how and when she takes action to meet the goal. The second imperative, in contrast, requests an action explicitly. Following Lorenzen [18, p.45], we call the first case *final imperatives*, and the second *a-final imperatives*. Finally, we allow for a corresponding *speech act*, a *request*, which expresses that someone is performing a request using an *imperative*.

³ “Aufforderung zur Herbeifuehrung eines Sachverhaltes”, see [18, p. 44]

► **Grammatical rule 13.** *request* : (person-nominator)? performance “(” imperative “)”

For example, Nora’s mother Ellie requests Nora to run some errands later:

“Ellie now π (Nora ! today do run this errand)”,

stands for the corresponding request. If we leave away the person nominators in such acts, we mean that the person who utters the request is requesting something from herself, meaning the person *sets herself a goal*. For example, I might now set myself the goal of running errands later today:

“now π (! today do run this errand)”

3 A pragmatic logic of experiments

In this section, we explain how the pragmatic language developed so far can be used to construct *logic formulas*, expressing experimental knowledge. Formulas can be used to express *experimental norms* for persons who should do something to perform an experiment, more specifically (and recursively), who should make claims, decisions and plans. To formalize experimental control, we introduce *practical modalities*. Furthermore, we use *experiential rules* to express claims about experimental outcomes. Rules can be tested by experiments and represented by *knowledge bases* and information models.

3.1 Knowledge of action consequences, inferences and decisions

In pragmatic philosophy, knowledge is understood as a form of *know-how*, meaning it must be *actionable*: knowledge enables action, encompassing the skills necessary to achieve goals, articulate and pursue interests, and ultimately navigate life within a heterogeneous society [17, 12]. What distinguishes knowledge from mere opinion is the notion of *validity*: a valid claim is a proposition that is successfully justifiable, which in turn requires the success of the actions underlying its defense, including the successful execution of experiments.

To be valid, claims must be generalizable across multiple examples. To express such generalizable claims, we employ standard logical connectives: *disjunction* (\vee) for “or,” *conjunction* (\wedge) for “and,” and *negation* (\neg) for “it is not the case that.” These can be combined to form complex propositions. Additionally, we use the *implication* operator (\rightarrow) to denote conditional statements: “if the first proposition is true, then the second must also be true.” For example, the logical formula $A \vee (\neg B \wedge (\neg(C \rightarrow D)))$ expresses a structured claim where A, B, C , and D are arbitrary propositions.

Quantifiers extend conjunction and disjunction over arbitrarily many propositions by introducing *variables*. Variables are placeholders for elements within a specified *domain* – a collection of nominators that share a common predicate. To denote domains, we use upper-case symbols corresponding to predicates in Sect. 2.1. For example, the domain *Person* consists of nominators referring to individuals. Variables such as x, y, z can be substituted by any element from their respective domains. The *universal quantifier* (\bigwedge) generalizes conjunction across all elements of a domain, asserting that a proposition holds for every substitution:

$$\bigwedge_{x \in \text{Space}} x \text{ now is raining} \wedge x \text{ now is wet}$$

This states that it is raining and wet everywhere in space. Conversely, the *existential quantifier* (\bigvee) generalizes disjunction, asserting that a proposition holds for at least one substitution:

$$\bigwedge_{x \in \text{Space}} \bigvee_{y \in \text{Time}} x \text{ } y \text{ is raining} \wedge x \text{ } y \text{ is wet}$$

This expresses that at every location in space, there exists some point in time where it is raining and wet. We refer to such quantified logical expressions as *formulas*. Formulas can be used to describe complex situations involving actions or processes.

A crucial aspect of pragmatic knowledge is understanding how actions lead to consequences. We distinguish between *conditions*, which must hold at the time an action is performed, and *consequences*, which describe the expected results. An action is deemed *unsuccessful* with respect to a goal if its consequences do not fulfill that goal. The reason for failure can often be traced back to unmet conditions. This leads to the notion of *knowledge about the consequences of actions*⁴. Such knowledge is formalized using *consequential rules*, which capture the expected outcomes of actions under specific conditions:

► **Schema 1** (consequential rules).

$$\bigwedge_{x, \dots, y \in D} (R(x, \dots, y) \wedge (\text{person-nominator})? \text{action}(x, \dots, y) \rightarrow EC(x, \dots, y))$$

Here, $R(x, \dots, y)$ denotes a formula capturing *requirements* (conditions necessary for the action), and $EC(x, \dots, y)$ denotes a formula capturing the *expected consequences*. For example:

$$\bigwedge_{x \in \text{Candle}} \bigwedge_{y \in \text{Matches}} \text{Nora now uses } y \text{ on } x \rightarrow x \text{ then is burning.}$$

This rule asserts that lighting a candle with a match under Nora's agency results in the candle burning – though this claim is context-dependent. It holds for an adult on Earth but fails for a child or in a zero-oxygen environment. In pragmatics, this only demonstrates the need to *refine* requirements for assuring validity. *Progression rules* describe changes in state over time due to processes rather than actions:

► **Schema 2** (progression rules).

$$\bigwedge_{x, \dots, y \in D} (R(x, \dots, y) \wedge (\text{endurant})? \text{happening}(x, \dots, y) \rightarrow EC(x, \dots, y))$$

The temporal ordering implicitly assumes that the antecedent conditions occur before the consequent state. For example:

$$\bigwedge_{x \in \text{Lake}} x \text{ now contains this amount of water} \wedge \text{here now raining that amount of water} \rightarrow x \text{ then contains (this + that) amount of water.}$$

Progression rules need to be justified by experiments (see below) or derived from other knowledge. A set of such rules forms a *rule base*: CRB for consequential rules and PRB for progression rules. Together with a set of formulas describing the current situation $S(t)$, we obtain a *knowledge base*: $CKB_{S(t)} = CRB \cup S(t)$ or $PKB_{S(t)} = PRB \cup S(t)$. If we can infer a formula F from such a knowledge base using logical inference, we write $KB \prec F$.

In addition to knowledge about consequences of actions and progressions, we also require knowledge about people's *behavior in terms of speech acts*. These are actions like *claims* and *requests* in which some explicit knowledge base is required. Correspondingly, we introduce *rules of inference* (for actions that derive claims from other claims) as well as *decision rules* (for deriving goals from other claims or other goals):

⁴ “Handlungsfolgenwissen” [12, 9]

► **Schema 3** (rules of inference). $\bigwedge_{o \in Person} (o \ t \ \pi(KB) \wedge o \ t \ \pi \ infer \rightarrow o \ (t + \delta) \pi(KB'))$

► **Schema 4** (decision rules). $\bigwedge_{o \in Person} (o \ t \ \pi(KB) \wedge o \ t \ \pi(! \Vdash S_g) \wedge o \ t \ \pi \ decide \rightarrow o(t + \delta) \pi(! \Vdash S_p))$

A particularly relevant example of a decision is to *plan*. Pragmatically, plans are understood as artifacts that are a result of a process of planning [8]. However, they are more than that: Plans are also symbolic manifestations of imperatives (formalized by using a request $\pi(!)$). For one, we plan according to a *planning goal*, which can be understood as a final imperative specifying an intended situation that should be realized by a plan. The plan itself manifests likewise a *final or an a-final imperative*, consisting of a series of actions to be performed or of subgoals to be pursued in order to reach this goal. A successful plan, thus, satisfies a conditioned imperative: it needs to successfully realize the goal whenever we follow it in an experiment. We can express this kind of knowledge also in terms of rules.

3.2 Practical modalities

Based on such knowledge bases, we can assess *what can be done*. Namely in the sense of knowing whether an *expected consequence A is achievable* in a given situation. The latter can be defined based on whether A is *logical implied* by consequential rules in this situation:

► **Definition 1** (A is achievable). $\Delta_{CKB_{S(t)}}^\pi A \leftrightarrow CKB_{S(t)} \prec A$

Literally, $\Delta_{CKB_{S(t)}}^\pi A$, or *A is achievable* means that some expected consequence described by the formula A can be justified by (repeatedly) applying consequential rules from the knowledge base to the situation $S(t)$. When it is clear which knowledge base is meant, we can also leave away the subscript: $\Delta^\pi A$.

The power of this *practical modal logic* [17, 18] is to capture everyday notions of *dispositions* and *action potentials* relative to a situation. This becomes clear when we define the modal variants:

► **Definition 2** (A is avoidable). $\overline{\Delta}^\pi A \leftrightarrow \Delta^\pi \neg A$

► **Definition 3** (A is unachievable). $\underline{\Delta}^\pi A \leftrightarrow \neg \Delta^\pi A$

► **Definition 4** (A is unavoidable). $\nabla^\pi A \leftrightarrow \neg \Delta^\pi \neg A$

► **Definition 5** (A is controllable). $\boxtimes^\pi A \leftrightarrow \Delta^\pi A \wedge \overline{\Delta}^\pi A$

If a consequence is avoidable, this means its contrary can be achieved. If it is unachievable, we fail to justify it can be achieved. And if it is unavoidable, we fail to justify that it can be avoided. For example, in a situation where a state launches atomic missiles to attack another state, which also possesses atomic missiles, an atomic war is unavoidable. This is because, according to our *knowledge of consequential rules of warfare* and assuming a certain *behavior*, namely that the corresponding protocols are implemented by the group of people responsible for them, we fail to find a path of action that would *not* involve launching a counter-attack, and thus we may not find a way to prevent a war in this situation.

Controllable situations are both achievable and avoidable. Sometimes we can avoid a consequence only constructively, based on changing a situation described in a corresponding formula using *another nominator*, i.e., to switch nominators. This leads to a more specific case of *value controllability*:

► **Definition 6** (A is (constructively) avoidable). $\bigwedge_D x. \overline{\Delta}_x^\pi A(x) \leftrightarrow \overline{\Delta}^\pi A(x) \wedge \bigvee_D x'. A(x')$

► **Definition 7** (A is (value) controllable). $\Sigma_x^\pi A(x) \leftrightarrow \Delta^\pi A(x) \wedge \overline{\Delta}_x^\pi A(x)$

The atomic counter-attack is a case in point, because there needs to be a switch for controlling the missile launch, and this switch is always in some position.

In an equivalent way, we can use modal logic to reason with knowledge of a situation and some *progression model*, which can be expressed as a collection of *progression rules*:

► **Definition 8** (necessary). $\nabla_{PKB_{S(t)}} A(t + \delta) \leftrightarrow PKB_{S(t)} \prec A(t + \delta)$

Literally, A is a *necessary consequence* of a given situation $S(t)$ at time $t + \delta$, under the assumption that the progression rules and the situation descriptions are defendable, and if $A(t + \delta)$ is a logical implication. By abstracting from the particular base $PKB_{S(t)}$, we also write $\nabla A(t)$ for the situations that will happen as a consequence of this situation at some time t . For example, in case we have a progression model of rainfall covering the extent of a lake, we may be able to predict the amount of water of that lake at a time after the rainfall stopped, given that we know its water content in the current situation. The definitions of these so called *mellontic modalities* [17] are equivalent to the practical ones above, including *possible* (Δ), *impossible* ($\overline{\Delta}$), and *contingent* (Σ). Contingent consequences are those that are possible yet we still fail to show that they are logically implied. That is, based on our progression model, *we just don't know*.

3.3 Experiments

The empirical (a-posteriori) knowledge [17] that we can obtain from an experiment can be written down in the form of *experiential rules* that are very similar to consequential rules introduced above, except that they involve the triggering of a *process* p (grammatical rule 4):

► **Schema 5** (experiential rule).

$$\bigwedge_{o \in Person} \bigwedge_{t \in Time} (S_c \wedge (o \text{ } t \text{ } \pi \text{ } (! \Vdash t \text{ } \kappa \text{ } p))) \rightarrow S_m$$

Literally, if we do something to start process p under the condition S_c , then we can expect situation S_m to occur [16]. Knowledge obtained from a given experiment can include many such rules. Experiential rules constitute both constructive building blocks and tests for empirical theories. In the latter case, by using a theory to infer an experiential rule that is compared with the result of a corresponding experiment, in the former case, by directly generalizing from experiential rules.

Yet, like all actions, experiments can *fail*, and in consequence, rules become invalid. How exactly can experiments fail? This depends on their purpose [10, 13]. The purpose of an experiment [16] derives from the trans-subjectivity of empirical knowledge: it is to *reproduce the process p such that it leads to similar situations (consequences) under the same conditions*, regardless of who is triggering the process and with which instruments (under which further circumstances). Conditions can be either *fixed* (not changed in the experiment) or *controlled* (changed in the experiment). This means that all conditions must be *achievable* via actions (definition 3), while controls need to be, in addition, *controllable* (can be switched on or off) (definition 7). In addition, we often need to leave some other situations *contingent* (“free”, or not pre-determined) (section 3.2, last paragraph). Conditions and contingent situations are required to prevent the experiment from being *disturbed*. The situations (grammatical rule 8) that are the consequences (schema 1) of the experiment can be represented by *measures*. Altogether, we call this the *experimental reproducibility norm*, and it has the following general form:

► **Schema 6** (experimental reproducibility norm).

An experiment $(F_1, \dots, F_k, C_1, \dots, C_n, p, M_1, \dots, M_u)$ is successful if the fixed conditions are achievable in situation S_f , the controlled conditions are controllable (in S_c) for each particular value c_1, \dots, c_n , and the situation S_u is contingent, and if, when achieving all conditions and starting the process p under arbitrary circumstances s_1, \dots, s_v , equivalent outcomes m_1, \dots, m_u occur (under some equivalence \equiv) in the resulting situation (S_m):

$$\begin{aligned} & \bigvee_{f_1, \dots, f_k \in F} \bigvee_{c_1, \dots, c_n \in C} \bigvee_{m_1, \dots, m_u \in M} \bigwedge_{s_1, \dots, s_v \in D'} \\ & \Delta^\pi S_f(f_1, \dots, f_n) \wedge \bigwedge_{c_1}^\pi S_c(c_1) \wedge \dots \wedge \bigwedge_{c_n}^\pi S_c(c_n) \wedge \bigwedge_{u \in D} u_1, \dots, u_m \cdot S_u(u_1, \dots, u_m) \wedge \\ & (S_f(f_1, \dots, f_n) \wedge S_c(c_1) \wedge \dots \wedge S_c(c_n) \wedge ((s_1, \dots, s_v) \pi(! \Vdash \kappa p) \rightarrow \\ & \bigvee_{m \in M} m'_1, \dots, m'_u \cdot S_m(m'_1, \dots, m'_u) \wedge m'_1 \equiv m_1, \dots, m_u \equiv m'_u) \end{aligned}$$

The nominators f_1, \dots, f_n (fixes, taken from domains F_i), c_1, \dots, c_n (controls, taken from domains C_i), u_1, \dots, u_m (contingents from domains U_i), m_1, \dots, m_u (measures, taken from domains M_i) thereby serve to identify and reproduce the respective situations.

For example, an experimental norm for a simple spatio-temporal experiment about growing crops in a geographic region could look like this:

► **Norm 1.** $\bigvee_{r \in \text{Region}} \bigvee_{m \in \text{AmountofBeans}} \bigwedge_{o \in \text{Person}} \bigwedge_{t \in \text{Time}}$
 $\bigwedge_r^\pi o \ t \ \pi \text{ sowing beans in } r \ \wedge \bigwedge_{u \in \text{AmountofBeans}} u.o \ (t + \delta) \ \pi \text{ selling } u \ \wedge$
 $((o \ t \ \pi \text{ sowing beans in } r \ \wedge o \ t \ \pi \text{ farming } \Vdash r \ t \ \kappa \text{ growing beans}) \rightarrow$
 $\bigvee_{m \in \text{AmountofBeans}} m'.o \ (t + \delta) \ \pi \text{ producing } m' \wedge m' \equiv m)$

This norm defines an experiment (*Region, grow beans, AmountofBeans*) to determine how many beans can be produced in a region r , independent of who performs it (o) or when (t). The experiment requires sowing beans in r at t (*controllable situation* S_c) and ensuring that later sales ($t + \delta$) do not interfere, avoiding market disturbances. If beans are sown and properly cultivated ($p = \text{grow}$), the norm expects that by ($t + \delta$), an approximate amount m of beans will be produced. This norm is *a priori*: it does not specify the exact yield but requires that outcomes be reproducible up to equivalence. Experiments implementing this norm either fix or control or leave contingent conditions when triggering the process. If reproducibility fails – e.g., due to lack of seeds, planting restrictions, or market constraints – the experiment fails.

In case of failure, we can adjust an experimental norm to ensure valid experiential rules. Lange [16] suggested the following principle ways to deal with such disturbances:

1. *Isolating* disturbances through shielding (possible in labs or simulations).
2. *Cleaning up* disturbances by controlling, fixing, or rendering them contingent (e.g., via randomization).
3. *Incorporating* disturbances as *errors*, increasing the tolerance of equivalences.

These adjustments constitute what Lange calls *fault avoidance knowledge* (referred to as *exhaustion* in [16]). For example, if bean growth depends on weather conditions or market quotas, fixing the yearly weather conditions and removing quota constraints could make the experiment reproducible. Note that *inferential statistics*, at its core, is a method for incorporating the disturbances of repeatable experiments using stochastic models (i.e., random generators) [18]. Methodologically, it comes *after* the introduction of experiments, not before.

Causal experiments play an exceptional role for science, since they allow us to determine *causes*. Yet, distinguishing causes from other experimental relations likewise requires pragmatic knowledge, an insight gained early by Georg Hendrik von Wright [28] in terms of his *interventionist causality norm*, and much later picked up in contemporary causal inference theory [20]. The corresponding experimental norm for causal experiments is more strict as it

requires in addition a particular *counterfactual* situation, i.e., considering a consequential situation that occurs if we had not taken an action [21]. The norm requires that if *some controls are not achieved*, then the corresponding *measures need to be different*:

► **Schema 7** (interventionist causality norm).

$$(S_f(f_1, \dots, f_n) \wedge \neg(S_c(c_1) \wedge \dots S_c(c_n)) \wedge ((s_1, \dots, s_v) \pi(! \Vdash \kappa p)) \rightarrow \neg \bigvee_{m \in M} m'_1, \dots, m'_u. S_m(m'_1, \dots, m'_n) \wedge m'_1 \equiv m_1, \dots, m_u \equiv m'_u)$$

If an experiment satisfies such a norm, there is a one-to-one correspondence between possible control situations and measure situations. This is the case, e.g., when we run *randomized control trials*, where a control group lacks the condition, and the experiment is successful in case that group also lacks the expected consequence [21]. We can then call the control domain a *cause* of the measure domain. In case of failure to satisfy such a norm, we can clean up disturbances, i.e., by incorporating conditions, or by adding contingencies into the norm. The corresponding strategies are well known from the causal reasoning literature [20], including fixing *confounders* (common causes of conditions and consequences), and leaving contingent *intermediators* (effects of controls that are causes of consequences) and *colliders* (common effects of controls and consequences) [21].

Data record experiential rules in terms of the underlying nominators (in our bean growing example $(r_1, m_1), \dots, (r_k, m_k)$). Yet, such data records leave away many details needed to understand the underlying experiment. This includes not only the irrelevant further circumstances (here: time and person), but in particular, the fact that fixes and controls uniquely determine (are *keys* for) measures, and the question what kind of situations are controlled, fixed, or measured. To keep some of this information in an abbreviated form, we use the following notation for the type of *experiential knowledge base* that corresponds to an experimental norm:

► **Definition 9** (experiential knowledge base).

$$EKB(f : X, c : Y, p : Process \rightarrow m : Z), \quad \text{where}$$

$$X, Y, Z = \begin{cases} D, & \text{domains of situation variables in an experiment} \\ \pi(KB), & \text{knowledge claims in an experiment} \\ \pi(! \Vdash \pi(KB)), & \text{requests for bringing about situations for knowledge claims} \end{cases}$$

Thus, for experiments, we usually control (c), fix (f) or measure (m) some *domains* D . For experiments that include *claims*, we additionally control, fix or measure knowledge claims ($\pi(KB)$) (which of course may be justified by further experiments). And for experiments that include *goals*, *decisions* and *plans*, we control, fix or measure requests for bringing about a situation in which we can make knowledge claims ($\pi(! \Vdash \pi(KB))$). For the fixed conditions, we also write down constants instead of the domain from which they stem.

4 Classes of spatio-temporal experiments

All other differentiation in experiments is a consequence of taking into account *different ways* of bringing about controls, triggering processes, and realizing measures [16]. An instrument for starting the process is called experimental apparatus. Instruments for observing and recording S_m are called measurement instruments. For measurements, we also need to control conditions, yet only for the process started within the sensor of the measurement instrument itself. An example for the latter would be a temperature measurement using a thermometer, where the process is the expansion of a thermometric material in the sensor [4], and among

17:12 Spatio-Temporal Experiments

the controlled conditions are, for example, the location and height above ground. A natural or “*quasi*” experiment is one in which the researcher does not control or fix the conditions of a process, but instead selects among conditions of processes that were already recorded. For *spatio-temporal experiments*, we distinguish the following classes, following Sinton [24], but enriched by more recent ideas about conceptual models of spatial information [23]. We specify experiments⁵ based on their underlying experimental norms:

► **Norm 2** (experimental norm for spatial fields).

$EKB(f : Time, c : Space, p : Process \rightarrow m : Endurance)$

Spatial fields fix time and control space in order to measure some endurance nominators (which could be amounts, stuff, objects). An example would be a raster map of forest density per grid cell.

► **Norm 3** (experimental norm for spatial coverages).

$EKB(f : Time, c : Endurance, p : Process \rightarrow m : AmountofSpace)$

Spatial coverages fix time and control endurances in order to measure some amount of space occupied by the endurance. An example would be a map of vector polygons of a land use, vegetation, or soil type.

► **Norm 4** (experimental norm for spatial lattices).

$EKB(f : Time, c : AmountofSpace, p : Process \rightarrow m : Endurance)$

Spatial lattices fix time and control an amount of space in order to measure some endurance controlled by this amount of space. An example would be statistical census tract data.

When using time as a control instead, we obtain various forms of *time series* experiments that involve space:

► **Norm 5** (experimental norm for temporal fields).

$EKB(f : Space, c : Time, p : Process \rightarrow m : Endurance)$

A *temporal field* controls time and fixes space, resulting in a time series that records measurements at a location over time. An example would be river discharge continuously measured at a catchment outlet, resulting in a hydrograph.

► **Norm 6** (experimental norm for trajectories).

$EKB(f : Endurance, c : Time, p : Process \rightarrow m : AmountofSpace)$

Trajectory experiments serve to measure motion, including movements of (rigid) objects (tracks) or spreadings etc. [6]. Note that spatio-temporal experiments are usually *not causal*, since they do not satisfy a counter-factual, interventionist causality norm. For example, when measuring a horizontal spatial temperature field, different locations will share the same temperature value, thus location cannot be considered a cause for temperature change. This is different when moving in the vertical direction (as temperature decreases with height). Yet, we can use causal experiments together with spatio-temporal measurements in order to infer knowledge in various ways, as illustrated in our example.

⁵ Note this is only a subset of possible spatio-temporal experiments.

5 The hidden experiments in landuse simulation modeling

In our sugarcane example case, we are interested in the question: what is the effect of one or more increased bioethanol demands on the spatial distribution of forest landuse [27]? With *hidden experiments*, we mean the (largely implicit) knowledge of the types of experiments that need to be mastered to answer this question. On the highest level of abstraction, our example corresponds to a *causal experiment*, where we need to control the bioethanol demand, *fix* conditions that also influence landuse (such as sugar demand), and *keep contingent* conditions that occur as *intermediators* of landuse planning goals, in order to infer a spatial distribution of landuse (forest) in a situation later $(t + \delta)$:

► **Norm 7** (Bioethanol demand landuse inference).

```
EKB(
  f :  $\pi(EKB(f : AmountofSpace, f : (t + \delta), p : demand \rightarrow m : AmountofSugar))$ ,
  c :  $\pi(EKB(f : AmountofSpace, f : (t + \delta), p : demand \rightarrow m : AmountofBioethanol))$ ,
  p : infer  $\rightarrow$ 
  m :  $\pi(EKB(f : (t + \delta), c : Landuse \rightarrow m : AmountofSpace))$ 
)
```

The problem is that the bioethanol demand needs to be *causally controlled*, meaning we need to compare the consequences of a demand increase with a reference scenario [27] in which the original demand remains the same, a scenario that has never been observed. Furthermore, landuse is subject to various invisible effects and human decisions that are not represented in observed landuse changes. Since we cannot actually control market demand, there is no way for us to *perform* a corresponding experiment. Furthermore, the problem can also not be solved by consulting past landuse images and running a *remote sensing experiment* over time: A remote sensing experiment controls locations or time and measures crop land type in terms of a field. Based on this, we can only measure landuse change over time and space in a non-causal manner, and only under the factual conditions of changing demands in the history of Brazil. It then becomes impossible to isolate the effects of bioethanol demand from sugar demand [27]. What we need instead is an experiment that measures the causal effects of invisible demands on decisions under counterfactual conditions.

For this reason, we need to construct a *model of the causal experiment*⁶, in which we can actively control the situations that trigger the process – such as in a simulation model. And for this purpose, we need to decompose the experiment into sub-experiments for which we can obtain some experiential knowledge to be used in the model. And here is where the task becomes really complex, because we have to figure out a way that these experiments feed into each other, see Fig. 3. First of all, the knowledge about the market demand needs to be input of a *decision experiment*. This experiment controls knowledge claims about the market demand at $t + \delta$ and produces a final plan with several subgoals, including the *sugarcane production goal* at $t + \delta$ for a certain spatial region. Here is a specification of the experimental norm:

► **Norm 8** (Sugarcane production decision).

```
EKB(
  f :  $\pi(EKB(f : AmountofSpace, f : (t + \delta), p : demand \rightarrow m : AmountofSugar))$ ,
  c :  $\pi(EKB(f : AmountofSpace, f : (t + \delta), p : demand \rightarrow m : AmountofBioethanol))$ ,
  p : decide  $\rightarrow$ 
  m :  $\pi(! \Vdash \pi(EKB(f : AmountofSpace, f : (t + \delta), p : produce \rightarrow m : AmountofSugarcane)))$ 
)
```

⁶ Cf. our definition in [23], where a model of an experiment is a method that answers the same question as the experiment.

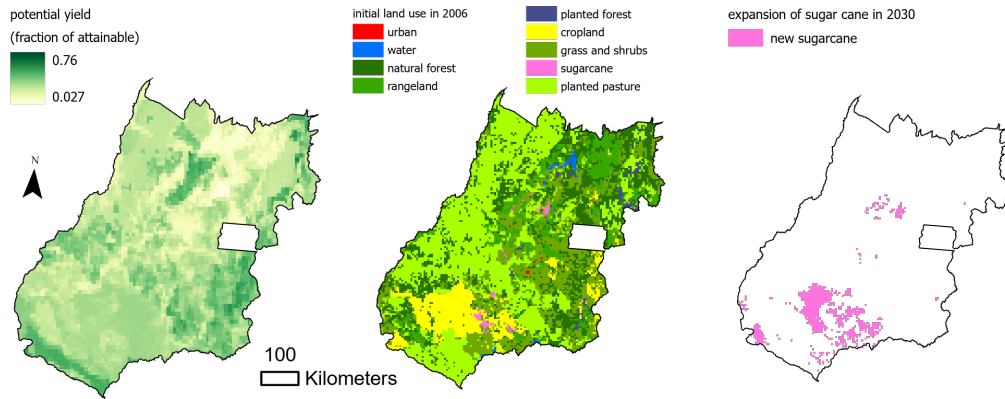
17:14 Spatio-Temporal Experiments

The result corresponds to a *lattice experiment*: for each region, we measure an amount of sugarcane that it should produce. The corresponding knowledge constitutes in turn a controllable input for a *planning experiment*, namely the decision of how to redistribute landuse to reach this production goal:

► Norm 9 (Landuse planning).

$$\begin{aligned}
 &EKB(\\
 &f : \pi(EKB(f : t, c : Space, p : grow \rightarrow m : AmountofSugarcane)), \\
 &f : \pi(EKB(f : t, c : Landuse \rightarrow m : AmountofSpace)), \\
 &c : \pi(! \vdash \pi(EKB(f : AmountofSpace, f : (t + \delta), p : produce \rightarrow m : AmountofSugarcane))), \\
 &p : plan \rightarrow m : \pi(! \vdash \pi(EKB(f : (t + \delta), c : Landuse \rightarrow m : AmountofSpace))) \\
 &)
 \end{aligned}$$

Note that in this planning experiment, the different production goals are competing because of a collider, which is the fixed total area available for landuse. Thus, if we increase sugar cane production, we need to decrease the production of other crops, pasture or forest. This is what demand-driven land use change models typically do, e.g. the models CLUE-S [26] and PLUC [27]. To perform this planning experiment, we need to fix claims about two kinds of further experiments, one is about the *sugarcane potential yield*, a *spatial field* that indicates for each location the potential sugarcane production density at the given time (t) (Figure 2).



■ **Figure 2** Potential yield of sugarcane (as fraction of the maximum attainable yield) (left), initial land use in 2006 (middle) and new locations with sugarcane cultivation in 2030 for a demand increase of 10.2 million m^3 ethanol, for the state Goiás in Brazil.

This knowledge, in turn, can be *obtained by inference* starting from a field of weather information and a field of soil types (the GAEZ method by the FAO)[27]:

► Norm 10 (Sugarcane yield inference).

$$\begin{aligned}
 &EKB(\\
 &f : \pi(EKB(f : t, c : Space, p : measure \rightarrow m : Soil)), \\
 &f : \pi(EKB(c : Time, c : Space, p : measure \rightarrow m : AmountofHeat)), p : infer \rightarrow \\
 &m : \pi(EKB(f : t, c : Space, p : grow \rightarrow m : AmountofSugarcane)) \\
 &)
 \end{aligned}$$

The second input condition is a knowledge claim about the current *landuse coverage* at time t (Figure 2), which can be obtained from remote sensing images. The planning experiment results in a single subgoal, namely the request to realize another landuse coverage at time $t + \delta$. The final step is to implement the plan and thus to realize the planned sugarcane production.

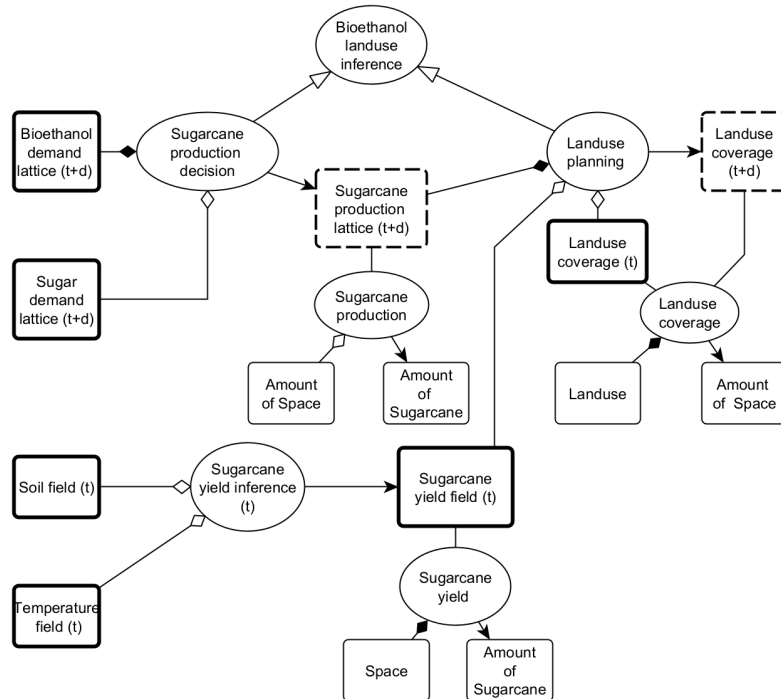


Figure 3 Experiments required for understanding the sugarcane example. Ellipses stand for experiments, round rectangles with thin borders denote domains, with thick borders knowledge bases, with dashed borders goals. Black diamonds are controls, white diamonds are fixed conditions. Black arrows denote measures. White arrows are sub-experiments.

6 Conclusion

In this paper, we proposed a formal pragmatic account of experiments to clarify their role in spatio-temporal modeling (Q). Our broader goal is to develop a systematic way to judge whether a given modeling approach is suitable for gaining knowledge about a particular type of experiment – especially those represented by spatial information models.

To this end, we introduced a grammar of situations and a pragmatic logic of experiments. This allows us to define experiments by their *experimental norms*, i.e., by distinguishing which experimental conditions must be fixed, controlled, or left contingent (via a practical modal logic), and by identifying the measured consequences as resulting from underlying actions that trigger processes (Q A). Causal experiments follow stricter, counterfactual norms. We then characterized experiential knowledge bases in terms of these norms, the domains of situation variables involved, the inferences made, and the goals pursued – particularly in contexts involving human decisions. Sinton’s structural ideas about spatio-temporal information were reframed in terms of non-causal experimental norms (Q B).

Using the sugarcane example, we showed how decomposing its components by experimental norms clarifies why remote sensing alone is insufficient to answer the question. We identified the need for additional experiments to assess indirect effects on deforestation, including decision-, planning-, and inference-experiments, as well as underlying spatio-temporal experiments – fields, lattices, and coverages (Q C).

This work lays the foundation for a theory that evaluates spatio-temporal models by their fitness for purpose (cf. [23]), independently of implementation details. Such a theory is urgently needed as machine learning models replace traditional approaches without accounting for purpose or experimental logic. Future work should expand the pragmatic logic across modeling examples, formalizing experiment decomposition and supporting reasoning about spatial designs and sampling strategies. In this sense, our work remains preliminary.

References

- 1 Marcos Adami, Bernardo Friedrich Theodor Rudorff, Ramon Morais Freitas, Daniel Alves Aguiar, Luciana Miura Sugawara, and Marcio Pupin Mello. Remote Sensing Time Series to Evaluate Direct Land Use Change of Recent Expanded Sugarcane Crop in Brazil. *Sustainability*, 4(4):574–585, 2012. doi:10.3390/su4040574.
- 2 Eugenio Y Arima, Peter Richards, Robert Walker, and Marcellus M Caldas. Statistical confirmation of indirect land use change in the Brazilian Amazon. *Environmental Research Letters*, 6(2):024010, 2011. doi:10.1088/1748-9326/6/2/024010.
- 3 Gilberto Camara, Max J Egenhofer, Karine Ferreira, Pedro Andrade, Gilberto Queiroz, Alber Sanchez, Jim Jones, and Lubia Vinhas. Fields as a generic data type for big spatial data. In *Geographic Information Science: 8th International Conference, GIScience 2014, Vienna, Austria, September 24-26, 2014. Proceedings 8*, pages 159–172. Springer, 2014. doi:10.1007/978-3-319-11593-1_11.
- 4 Hasok Chang. *Inventing temperature: Measurement and scientific progress*. Oxford University Press, 2004. doi:10.1093/0195171276.001.0001.
- 5 Nicholas Chrisman. *Exploring Geographic Information Systems, 2nd Edition*. Wiley, 2002.
- 6 Antony Galton. *Qualitative spatial change*. Oxford University Press, 2000.
- 7 Antony Galton. Fields and objects in space, time, and space-time. *Spatial cognition and computation*, 4(1):39–68, 2004. doi:10.1207/s15427633scc0401_4.
- 8 Armin Grunewald. Kulturalistische Planungstheorie. In *Methodischer Kulturalismus: Zwischen Naturalismus Und Postmoderne*, pages 315–343. Suhrkamp, 1996.
- 9 Dirk Hartmann. Kulturalistische Handlungstheorie. In *Methodischer Kulturalismus: Zwischen Naturalismus Und Postmoderne*, pages 70–114. Suhrkamp, 1996.
- 10 Peter Janich. Das Experiment in der Psychologie. In *Konstruktivismus und Naturerkenntnis*, pages 275–289. Suhrkamp, 1996.
- 11 Peter Janich. Methodical constructivism. In *Issues and Images in the Philosophy of Science*, pages 173–190. Springer, 1997.
- 12 Peter Janich. *Logisch-pragmatische Propaedeutik*. Velnbrueck Wissenschaft, 2001.
- 13 Peter Janich. Das Experiment in der Biologie. In *Kultur und Methode*, pages 330–366. Suhrkamp, 2006.
- 14 Krzysztof Janowicz, Song Gao, Grant McKenzie, Yingjie Hu, and Budhendra Bhaduri. GeoAI: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond. *International Journal of Geographical Information Science*, 34(4), 2019. doi:10.1080/13658816.2019.1684500.
- 15 Werner Kuhn. Core concepts of spatial information for transdisciplinary research. *International Journal of Geographical Information Science*, 26(12):2267–2276, 2012. doi:10.1080/13658816.2012.722637.

- 16 Rainer Lange. Vom Koennen zum Erkennen. Die Rolle des Experimentierens in den Wissenschaften. In *Methodischer Kulturalismus: Zwischen Naturalismus Und Postmoderne*, pages 157–196. Suhrkamp, 1996.
- 17 Paul Lorenzen. *Normative logic and ethics*. Bibliographisches Institut, Mannheim, 1969.
- 18 Paul Lorenzen. *Lehrbuch der konstruktiven Wissenschaftstheorie*. Springer, 1987.
- 19 John McCarthy and Patrick J Hayes. Some philosophical problems from the standpoint of artificial intelligence. In *Readings in artificial intelligence*, pages 431–450. Elsevier, 1981.
- 20 Judea Pearl. *Causality*. Cambridge university press, 2009.
- 21 Mattia Proserpi, Yi Guo, Matt Sperrin, James S Koopman, Jae S Min, Xing He, Shannan Rich, Mo Wang, Iain E Buchan, and Jiang Bian. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence*, 2(7):369–375, 2020. doi:10.1038/s42256-020-0197-y.
- 22 Simon Scheider and Kai-Florian Richter. Pragmatic geoai: Geographic information as externalized practice. *KI-Künstliche Intelligenz*, 37(1):17–31, 2023. doi:10.1007/s13218-022-00794-2.
- 23 Simon Scheider and Judith A Verstegen. What is a spatio-temporal model good for?: Validity as a function of purpose and the questions answered by a model. In *16th International Conference on Spatial Information Theory (COSIT 2024)*, pages 7–1. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2024. doi:10.4230/LIPIcs.COSIT.2024.7.
- 24 David Sinton. The inherent structure of information as a constraint to analysis: Mapped thematic data as a case study. *Harvard papers on geographic information systems*, 1978.
- 25 Eric Top, Simon Scheider, Haiqi Xu, Enkhbold Nyamsuren, and Niels Steenbergen. The semantics of extensive quantities within geographic information. *Applied Ontology*, 17(3):337–364, 2022. doi:10.3233/AO-220268.
- 26 Peter H. Verburg and Koen P. Overmars. Dynamic simulation of land-use change trajectories with the CLUE-s model. In Eric Koomen, John Stillwell, Aldrik Bakema, and Henk J. Scholten, editors, *Modelling land-use change: Progress and applications*, pages 321–337. The GeoJournal Library, 2007. doi:10.1007/978-1-4020-5648-2_18.
- 27 Judith A Verstegen, Floor van der Hilst, Geert Woltjer, Derek Karssenbergh, Steven M de Jong, and André PC Faaij. What can and can’t we say about indirect land-use change in Brazil using an integrated economic–land-use change model? *Gcb Bioenergy*, 8(3):561–578, 2016. doi:10.1111/gcbb.12270.
- 28 Georg Henrik Von Wright. *Explanation and understanding*. Cornell University Press, 2004.

U-Prithvi: Integrating a Foundation Model and U-Net for Enhanced Flood Inundation Mapping

Vit Kostejn 

Charles University, Prague, Czech Republic

Yamil Essus 

Industrial and Systems Engineering Department, North Carolina State University,
Raleigh, NC, USA

Jenna Abrahamson 

Center for Geospatial Analytics, North Carolina State University, Raleigh, NC, USA

Ranga Raju Vatsavai 

Computer Science Department, North Carolina State University, Raleigh, NC, USA

Abstract

In recent years, large pre-trained models, commonly referred to as foundation models, have become increasingly popular for various tasks leveraging transfer learning. This trend has expanded to remote sensing, where transformer-based foundation models such as Prithvi, msGFM, and SatSwinMAE have been utilized for a range of applications. While these transformer-based models, particularly the Prithvi model, exhibit strong generalization capabilities, they have limitations on capturing fine-grained details compared to convolutional neural network architectures like U-Net in segmentation tasks. In this paper, we propose a novel architecture, U-Prithvi, which combines the strengths of the Prithvi transformer with those of U-Net. We introduce a RandomHalfMaskLayer to ensure balanced learning from both models during training. Our approach is evaluated on the Sen1Floods11 dataset for flood inundation mapping, and experimental results demonstrate better performance of U-Prithvi over both individual models, achieving improved performance on out-of-sample data. While this principle is illustrated using the Prithvi model, it is easily adaptable to other foundation models.

2012 ACM Subject Classification Computing methodologies → Image segmentation

Keywords and phrases GeoAI, flood mapping, foundation model, U-Net, Prithvi

Digital Object Identifier 10.4230/LIPIcs.GIScience.2025.18

Supplementary Material *Software (Source Code):* https://github.com/kostejn/prithvi_segmentation, archived at `swh:1:dir:9af733737a3cc731b66abb19b15b204a01ed28c1`

Funding *Jenna Abrahamson:* Funded by a National Science Foundation Graduate Research Fellowship Grant No. DGE-2137100.

1 Introduction

Floods are one of Earth's most devastating natural disasters, and their impact is expected to intensify in a warmer climate. Variations in extreme temperatures and heavy rain events are expected to increase the frequency and intensity of floods, with implications for infrastructure stability, water quality, and human safety [16]. Understanding when and where floods occur is important not only for disaster response management, but also for understanding global hydrological and biogeochemical cycles [5]. Satellite remote sensing has been used since the 1970's to map surface water across the globe [15], including water from flood events. The public release of satellite data archives [36], combined with improvements in computing power and artificial intelligence algorithms, has led to large advancements in recent years for



© Vit Kostejn, Yamil Essus, Jenna Abrahamson, and Ranga Raju Vatsavai;
licensed under Creative Commons License CC-BY 4.0

13th International Conference on Geographic Information Science (GIScience 2025).

Editors: Katarzyna Sila-Nowicka, Antoni Moore, David O'Sullivan, Benjamin Adams, and Mark Gahegan;
Article No. 18; pp. 18:1–18:17



Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

mapping flood events across space and time [34]. Both optical and radar data have proven useful for capturing floods, and the amount of available public and commercial satellite data continues to grow in both of these domains.

This rapid increase in available data has led to the adoption of deep-learning techniques for flood mapping. Deep learning algorithms are inspired by the structure and function of brain neural networks, where the computer learns a hierarchy of concepts, each concept defined through its relationship to simpler concepts in a deep graph with many layers [11]. Deep learning for remote sensing of flood extents is commonly approached as a semantic segmentation task, using methods such as Convolutional Neural Networks (CNNs) like DeepLab [7], SegNet [9], and U-Net [29], a review of which can be found in [2]. Although deep learning has greatly improved the accuracy with which flood extents can be classified, it requires substantial amounts of labeled data, can be expensive to train, and does not always generalize globally across space and time, necessitating the training of application- or location-specific models. These shortcomings have spurred the recent development of novel geospatial foundation models. A foundation model is a large, generalizable deep learning model that is pre-trained on a massive, unlabeled dataset to be a generalist model that can then be fine-tuned to a variety of downstream tasks using smaller labeled datasets [17].

The applicability and accuracy of these foundation models compared to other deep learning methods are still being tested, especially in the geospatial field. Recently, NASA and IBM teamed up to release the novel Prithvi geospatial foundation model, which was pre-trained on data for the United States from the Harmonized Landsat-Sentinel (HLS) satellite data catalog. The HLS data catalog consists of harmonized data from both Landsat and Sentinel-2 satellite missions, with an average revisit period of cloud-free imagery every 8.4 days at 30 m spatial resolution [8]. Initial experiments of adapting the Prithvi model to flood mapping [19] compared the foundation model to a U-Net model using Sentinel-2 satellite imagery at 10 m spatial resolution. In this study, the authors of [19] found that the U-Net model outperformed the Prithvi model when evaluated on in-sample test data; however, Prithvi was found to perform better than the U-Net when evaluated on out-of-sample data from an unseen region in Bolivia. In their results, the authors show that both models performed poorly in at least one of the test sets (in-sample/out-of-sample). This finding suggests that a mixed approach could take advantage of both generalization capabilities and region-specific learning.

Building on Prithvi’s strong generalization capabilities, this paper proposes U-Prithvi, a novel fusion model combining Prithvi with U-Net for flood extent mapping in satellite imagery. This fusion approach aims to leverage cross-modal learning between the two models, capitalizing on Prithvi’s generalist strengths and U-Net’s detailed segmentation capabilities within local datasets. Using the Sen1Floods11 hand-delineated flood dataset, we: (1) trained a U-Net and Prithvi fusion architecture; (2) fine-tuned our proposed U-Prithvi model; and (3) fine-tuned and trained standalone U-Net and Prithvi models for performance comparison against our results and those reported in the literature. We evaluated these models against in-sample and out-of-sample test datasets from Sen1Floods11, assessing their flood extent mapping accuracy using Sentinel-2 imagery. Following an overview of each model and related work, we detail their architectures. Finally, we present our findings on how novel foundation models like Prithvi can be combined with established models like U-Net to improve global flood extent mapping accuracy.

The remainder of this paper is structured as follows. Section 2 presents relevant literature. Section 3 describes our models and performance evaluation methodology. Section 4 describes our results and summarizes our findings. Finally, Section 5 presents our conclusions.

2 Related Work

2.1 Deep Learning & Flood Mapping

Deep learning approaches have been implemented for a variety of remote sensing classification problems pertaining to land cover, agriculture, open water bodies, and floods. Floods present a unique problem in that they can be spectrally very complex depending on the geographic area and terrain that the flood is covering. Examples of deep learning in the literature for flood mapping include various applications of CNNs, Multi-Layer Perceptrons, and Recurrent Neural Networks applied to Sentinel-2, Sentinel-1, and CubeSat data [6, 22, 31], a full review of which can be found in [4]. Specific flood mapping architectures have also been developed, such as Siam-DWENet [38], which takes advantage of transfer learning and an attention mechanism, and DeepFlood, which employs feature-level fusion and classification of optical and radar data [18]. An architecture that is increasingly common for flood-mapping applications is U-Net, which has been used to classify floods based on satellite imagery and Twitter data [30], and for flood segmentation using radar data in southeastern Mexico [27].

2.2 U-Net Architecture

The U-Net architecture was first introduced by Ronneberger et. al 2015 [29] as an improvement of fully-convolutional neural networks for biomedical image segmentation applications. The U-Net consists of a series of contracting convolutional layers that gradually reduce spatial dimensionality and increase feature dimensionality. The feature-dense bottleneck is then upsampled with skip connections that help map from a condensed feature space using higher resolution information available in the encoder hidden layers.

The U-Net architecture has been extensively used for medical image segmentation [3]. However, models trained to perform natural image segmentation, like those in medical images or classic computer vision tasks, do not translate directly to remote sensing applications [33]. Recent advances in applying U-Net for remote sensing segmentation problems include combinations of DenseNet and Dilated Convolutions [33] and attention mechanisms [37], as well as specific applications to flood detection using SAR imagery [20].

2.2.1 U-Net Extensions

Although U-Net is highly effective in capturing both global and local context, it has limited capacity to learn long-range spatial dependencies [13]. This limitation has motivated research into integrating transformers with U-Net to enhance its performance. One notable architecture, UNETR, was proposed for 3D medical image segmentation and incorporates a Vision Transformer (ViT) with a U-shaped CNN structure. In this model, the transformer encoder is directly connected to the U-Net-style decoder through skip connections at different resolutions [13]. Other studies have modified U-Net by adding an additional transformer encoder, then combining its features with those of the CNN. This approach has been applied successfully in medical image segmentation with the FT-UNet architecture [35] and in remote sensing image segmentation with the ST-UNet architecture [14]. Petit et al. [28] introduced a U-Net variant with an attention mechanism for medical image segmentation. They incorporated self-attention into the bottleneck layer and cross-attention into the skip connections.

2.3 Foundation Models and Prithvi

Deep learning approaches require substantial amounts of ground truth data for training from scratch, which can be costly to obtain. This need for data, combined with the recent popularity of foundation models, large pre-trained models that have greatly impacted fields such as natural language processing and multimodal tasks, has driven research into foundation models specifically for the geospatial field. As evidence of this trend, research on transfer learning for geospatial tasks saw a ten-fold increase in published articles between 2017 and 2022 [21]. The use of foundation models in geospatial tasks can generally be divided into three categories: (1) models trained on natural image datasets, (2) models trained on geospatial datasets, and (3) hybrid models that integrate both approaches [23].

The first category typically involves models trained on datasets like ImageNet [10]. Although this dataset is quite different from geospatial datasets, studies have shown that this approach can still be effective for specific satellite image tasks, such as land-use classification, urban zone classification, or burnt area detection [25, 26]. In the second category, numerous foundation models have been developed specifically for remote sensing data. Some of these models are trained on single-time images but incorporate data from multiple remote sensors [12], while others are trained on multi-temporal datasets, such as the Prithvi model [17] and SatSwinMAE [24].

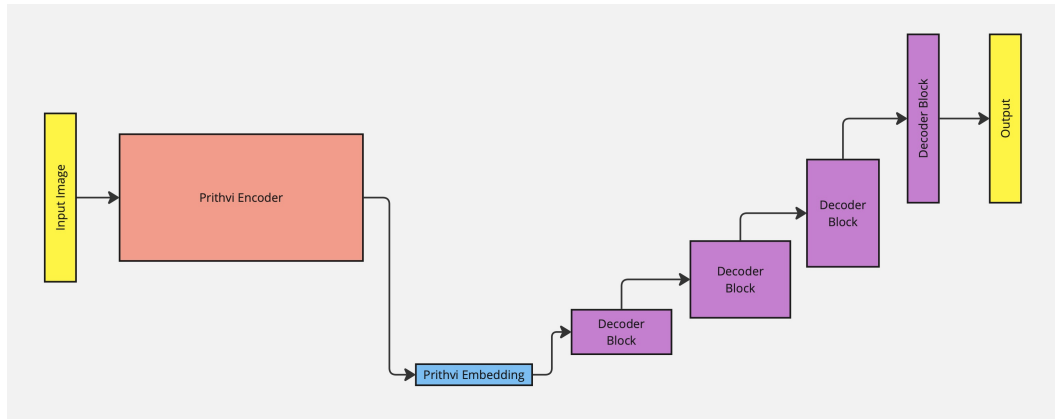
Several hybrid models combining natural and remote sensing image modalities have emerged in recent years. For instance, the GFM model [23] employs a two-stage pretraining process. It is initially trained on the ImageNet22k dataset and subsequently on the custom GeoPile remote sensing dataset. This approach aims to enhance performance compared to models trained solely on natural image datasets, while also mitigating the expense of training foundation models from scratch on remote sensing data.

2.3.1 Prithvi

Prithvi is a foundation model developed by IBM and NASA, specifically trained from scratch on geospatial data. Its encoder architecture is based on a Vision Transformer and was trained using a Masked AutoEncoder strategy. The training dataset for Prithvi Version 1 (the most recent version during the time of writing) comprises over 1 TB of multispectral satellite imagery from the HLS dataset, which was collected using a stratified sampling procedure to ensure a set of diversified data from the United States. As a multi-temporal model, Prithvi can process entire image sequences, allowing for time-series analysis of geospatial data [17].

2.4 Current Limitations and Our Contributions

Flood segmentation mapping using GeoAI foundation models has demonstrated effective generalization. However, these models often underperform on in-sample test datasets compared to U-Net models [19]. Let us formally define in-sample and out-of-sample test data. An in-sample test dataset consists of samples drawn from the same distribution as the training data (e.g., from the same regions). An out-of-sample test dataset contains samples from a different distribution but shares the same features (e.g., a different region), where the trained models have not seen any samples from that region. We hypothesize that combining Prithvi and U-Net will improve performance for both in-sample and out-of-sample use cases. Therefore, we developed a novel fusion model integrating Prithvi features into a U-Net architecture. We validated this hypothesis by comparing our model’s generalization and predictive capacity against existing approaches on both in-sample and out-of-sample test datasets. Our



■ **Figure 1** Our architecture for semantic segmentation that uses Prithvi as an encoder.

experiments show that this combined architecture balances these two performance metrics while requiring fewer training epochs. Finally, we discuss the potential extension of this approach to other foundation models and applications.

3 Methodology

This section describes the proposed approach, focusing on a novel architecture that combines the U-Net and Prithvi models. By merging U-Net’s ability to capture fine-grained details with Prithvi’s capacity to model global context, we aim to achieve superior segmentation accuracy for flood mapping. To evaluate this architecture, we will compare it against both models used independently. The following sections detail each model’s architecture before describing our fusion approach, U-Prithvi.

3.1 Prithvi Architecture

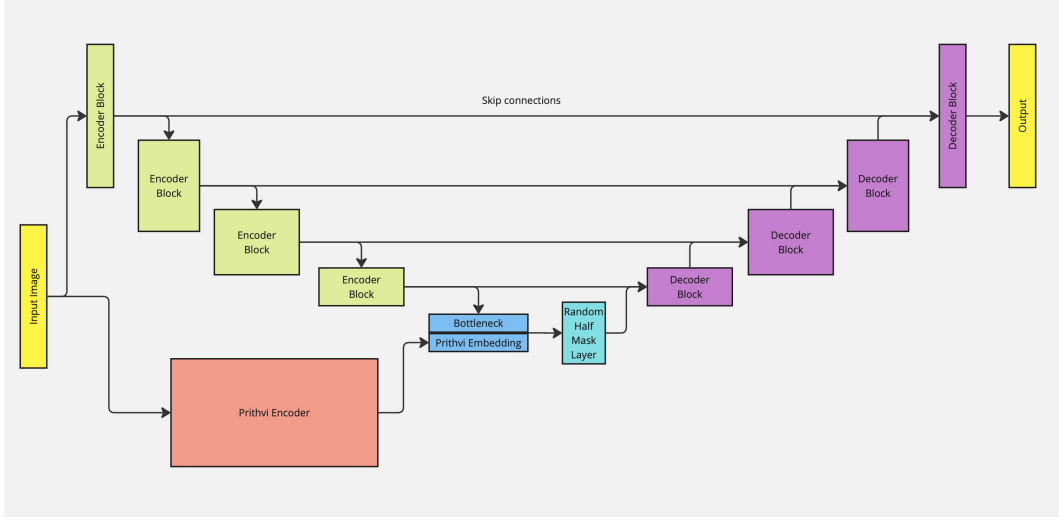
Prithvi is a Vision Transformer model trained using a Masked Autoencoder strategy. While its original decoder is designed for input reconstruction and may not be optimal for segmentation tasks [17], we utilize only Prithvi’s encoder and pair it with a custom decoder similar to that of U-Net.

Our implementation, illustrated in Figure 1, begins with Prithvi’s encoder, producing a 14x14 feature map with 768 filters. This feature map is processed through a single block without parameter modification, then upsampled using four transposed convolutional layers to achieve the original input shape. A final convolutional layer with Softmax activation serves as the classifier. Prithvi and U-Net serve as baselines in our experiments.

3.2 Proposed Fusion Model: U-Prithvi

U-Prithvi, the proposed model, integrates the strengths of both the Prithvi model and the U-Net architecture, combining Prithvi’s capability for capturing global context with U-Net’s proficiency in fine-detail segmentation. The schematic representation of the U-Prithvi architecture is illustrated in Figure 2.

The input passes through both the U-Net and Prithvi encoders, generating two 14x14 feature maps with 768 filters each. These feature maps are aggregated and subsequently passed into the decoder component, which comprises four upsampling blocks with skip connections to the U-Net encoder.



■ **Figure 2** Proposed architecture of U-Prithvi.

To address the potential training imbalance between the pretrained Prithvi encoder and the untrained U-Net encoder, we introduce a novel RandomHalfMaskLayer (RHM layer). This layer, positioned after the concatenation step, probabilistically masks either Prithvi’s or U-Net’s feature maps – or leaves both unmasked – during training. During inference, the layer has no effect on the input. We anticipate that this approach will promote balanced training and ensure optimal contributions from both components in the final segmentation.

Despite the existence of several approaches that combine transformer and U-Net architectures, most of these train the architecture from scratch and do not account for using fixed pretrained architectures. The Prithvi foundation model, being a ViT-based architecture, does not include multiple stages with varying resolutions, making it incompatible with the methods presented in the related work section 2.2.1. A significant advantage of our approach is its flexibility: Prithvi can be substituted with any other foundation model without requiring substantial modifications to the architecture.

4 Experimental Design and Results

Our experiments explored the following key research questions:

1. Do foundation models like Prithvi generalize well to out-of-sample data, while custom models like U-Net perform better on in-sample data?
2. Can we combine Prithvi and U-Net to exploit cross-modal relationships and improve performance on both in-sample and out-of-sample test datasets?
3. Can the U-Net architecture be improved to match Prithvi’s out-of-sample performance?
4. Can cross-modal learning be controlled?

We also conducted ablation experiments to investigate fusion strategies.

4.1 Data and Performance Metrics

Using the Sen1Floods11 [1] dataset, we evaluated the performance of our U-Prithvi architecture and compared it against standalone U-Net and Prithvi-decoder models from [19]. The Sen1Floods11 dataset contains 446 image samples paired with hand-labeled masks identifying

■ **Table 1** Performance metrics of the U-Net and Prithvi from [19] and our implementation of the Prithvi-decoder architecture as well as our proposed model U-Prithvi. All values represent percentages. “Test” label is used to denote the in-sample test set, while “Bolivia” denotes out-of-sample test set.

Model	Data Set	IoU			Accuracy		
		Avg.	Flood	Non-Flood	Avg.	Flood	Non-Flood
U-Net _{Base} [19]	Test	90.80	84.03	97.57	94.80	90.74	98.86
Prithvi _{Base} [19]	Test	89.59	81.98	97.21	94.35	90.12	98.58
Prithvi	Test	86.25	76.39	96.10	93.90	90.40	97.40
U-Prithvi	Test	89.73	82.21	97.24	94.81	91.15	98.46
U-Net _{Base} [19]	Bolivia	82.54	70.57	94.52	86.45	73.73	99.18
Prithvi _{Base} [19]	Bolivia	86.02	76.62	95.43	90.38	82.12	98.65
Prithvi	Bolivia	82.89	72.42	93.36	93.24	91.61	94.88
U-Prithvi	Bolivia	87.70	79.68	95.71	93.31	88.84	97.78

flooded (water and flood) and non-flooded areas across various regions, including Ghana, India, the Mekong River, Nigeria, Pakistan, Paraguay, Somalia, Spain, Sri Lanka, and the USA. Each sample has a resolution of 10 meters and a pixel size of 512×512 . Six Sentinel-2 bands (RGB, NIR, SWIR1-2) are used as input to our models to align with the Prithvi encoder’s input requirements.

Our preprocessing pipeline includes data normalization (mean 0, variance 1), random cropping to 224×224 (to meet Prithvi encoder’s input requirements), and random horizontal and vertical flips to augment the data. For performance evaluation, we use two sets provided in the dataset: (1) an in-sample test set with samples from the same regions as the training data, and (2) an out-of-sample test set containing samples from Bolivia, which were not used during training. To compare model performance, we use Intersection over Union (IoU) as our primary metric and accuracy (Acc) as a secondary metric. For each metric, we calculate both the mean (mIoU, mAcc) across classes and individual values for each class ($\text{IoU}_{\text{flood}}$, $\text{IoU}_{\text{nonflood}}$, $\text{Acc}_{\text{flood}}$, $\text{Acc}_{\text{nonflood}}$). These values are derived from true positive (TP), false positive (FP), true negative (TN), and false negative (FN) counts as follows:

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}} \quad (1)$$

$$\text{mIoU} = \frac{\text{IoU}_{\text{flood}} + \text{IoU}_{\text{nonflood}}}{2} \quad (2)$$

$$\text{Acc} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

$$\text{mAcc} = \frac{\text{Acc}_{\text{flood}} + \text{Acc}_{\text{nonflood}}}{2} \quad (4)$$

4.2 Results

The first experiment is designed to answer our first research question. The baseline performance of Prithvi versus U-Net for this application of flood inundation mapping was established in the paper published by [19]. We utilized the same in-sample and out-of-sample test sets as well as computed accuracy metrics to maintain consistency and comparability. These results show that U-Net_{Base} outperforms Prithvi_{Base} across all accuracy metrics when evaluated on in-sample test set as shown in Table 1. However, Prithvi_{Base} outperforms U-Net_{Base} when evaluated on out-of-sample Bolivia set. These results will be used as our baseline to compare to when analyzing the results of our experiments.

Our second experiment is designed to address our second and most important research question. By integrating the U-Net and Prithvi models, we anticipate that the U-Prithvi model will perform effectively on both same-distribution and unseen data, while demonstrating fast convergence for both scenarios. Table 1 presents our performance results compared to U-Net_{Base} and Prithvi_{Base} , evaluated on both the in-sample test set and the out-of-sample Bolivia dataset. Our findings indicate that the U-Prithvi model outperforms both approaches on out-of-sample data. For in-sample data, the performance of U-Prithvi falls between the two models, with U-Net dominating in this context. This suggests that the U-Prithvi architecture combines the ability to achieve strong predictive performance on datasets similar to the training set with the capacity to leverage the pre-trained foundation model for superior generalization on unseen data, without requiring additional training cycles.

4.2.1 Qualitative analysis

In Figure 3, we present a comparison of U-Prithvi with each of the competing models for three testing instances. In each figure, the left pane represents the satellite input, the middle pane a comparison between U-Net and U-Prithvi, and the right pane a comparison between Prithvi and U-Prithvi classification outputs. In each classification pane, we color-coded the pixels to highlight relative performance. In particular, green pixels means correctly classified by U-Prithvi and incorrectly classified by the competing model, red pixels are the opposite and blue pixels represent incorrect classifications for both. Figures 3a and 3b show test samples from the same-distribution dataset while Figure 3c shows an example of the Bolivia dataset.

We find that in most situations, U-Prithvi shows better performance in fine-detail areas, such as the borders of the floods in Figures 3a and 3c. This is especially noticeable when compared to the Prithvi model. Additionally, Figure 3b shows an example where U-Prithvi correctly identified entire non-flooded regions that the Prithvi model incorrectly classified. The ability to improve over the Prithvi predictions in areas with fine spatial details while still performing well against the U-Net model is one of the advantages of using this fusion architecture.

Some features are still difficult for the models to capture. For instance, none of the models correctly classified the river and some inland areas in figure 3a. Finally, Table 1 includes our implementation of the Prithvi model. Because [19] does not report the exact parameters of their model, we are not able to reproduce the model. Our implementation of the Prithvi-decoder architecture shows weaker performance than that reported in [19].

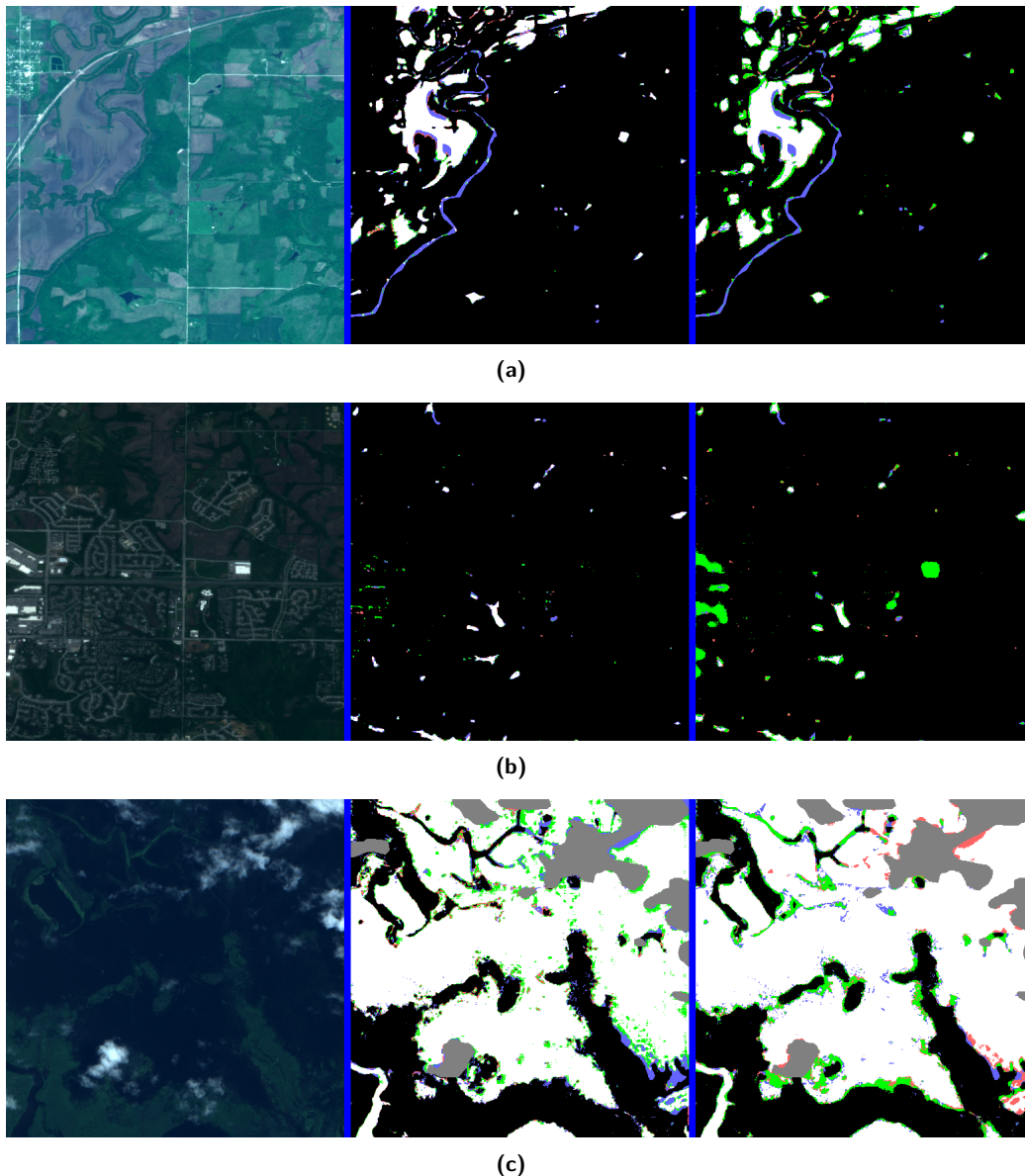
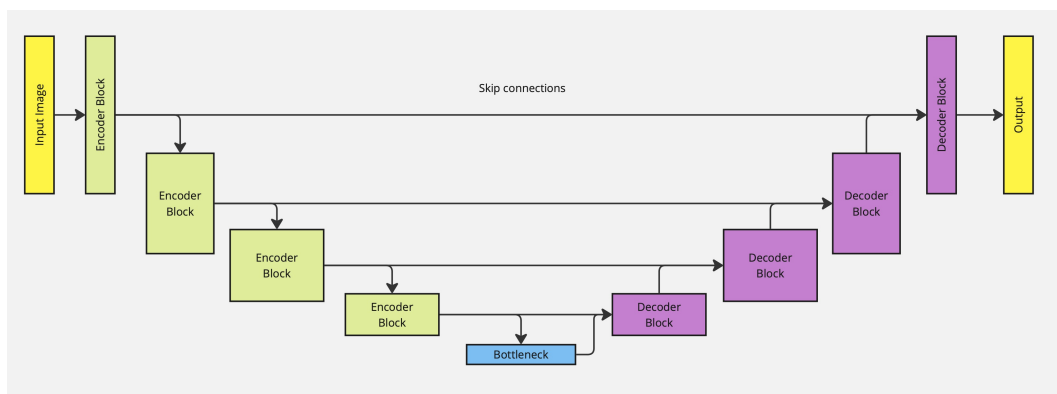


Figure 3 Classification outputs for different satellite inputs. Left pane is satellite input, middle is U-Net and U-Prithvi predictions and right is Prithvi and U-Prithvi predictions. A green pixel represents a pixel correctly classified by U-Prithvi and incorrectly classified by the other model. A red pixel means the pixel was incorrectly classified by U-Prithvi but correctly classified by the other model. A blue pixel represents incorrect classification for both. Finally, black and white pixels are correct classifications by both models of non-flooded and flooded pixels respectively. Grey pixels represent input labeled unclear.

4.3 Improving the U-Net Architecture

Our third research question is addressed in this section. The U-Net architecture is widely used in semantic segmentation [32] and is less computationally and data-intensive to train compared to U-Prithvi. Because of this, we expanded the architecture from [19] and found that increasing the size of the U-Net model can yield similar performance for the benchmark dataset.



■ **Figure 4** U-net architecture used in this paper.

■ **Table 2** Performance metrics of the revised U-Net model against the best-performing U-Prithvi. The U-Net model was trained for 100 epochs and U-Prithvi for 50+50 epochs. All values represent percentages.

Model	Data Set	IoU			Accuracy		
		Avg.	Flood	Non-Flood	Avg.	Flood	Non-Flood
U-Net _{ours}	Test	88.84	80.69	96.99	94.21	90.06	98.35
U-Prithvi	Test	89.73	82.21	97.24	94.81	91.15	98.46
U-Net _{ours}	Bolivia	87.77	79.88	95.65	94.1	90.85	97.34
U-Prithvi	Bolivia	87.70	79.68	95.71	93.31	88.84	97.78

Our U-Net implementation, shown in Figure 4, comprises three main components: Encoder, Bottleneck, and Decoder. Each component includes several blocks, each consisting of a convolutional layer, Batch Normalization, and a ReLU activation function. The stride and type of convolution (normal or transposed) determine whether the block reduces, maintains, or increases feature map resolution.

The encoder downscales the input into a compact feature representation, using four downsampling blocks (stride 2) that reduce the resolution from 224 to 14 while increasing filter depth from 6 bands to 768. A single bottleneck block preserves the resolution and feature count. The decoder then upscales the feature map back to the input image resolution through four upsampling blocks, each concatenated with the corresponding encoder block. A final convolutional layer provides the desired output shape. This yields a total of over 37 million trainable parameters, compared to the 29 million reported in [19].

The performance of our revised U-Net model is presented in Table 2 alongside the performance of U-Prithvi in both datasets. Our results show that for the Sen1Floods11 dataset, it is possible to achieve results comparable to the performance of U-Prithvi using the U-Net model with a higher count of parameters. This is true for both in-sample and out-of-sample test sets. In the case of the Bolivia dataset, we find that performance is practically identical compared to the U-Prithvi model. This implies that a larger model is still able to capture the dynamics of the data, even compared to a foundation model. Nevertheless, it is unclear if this result would also be evidenced using a different dataset.

■ **Table 3** Performance metrics of the U-Prithvi model trained with and without the RHM layer. All values represent percentages and all configurations were trained using 50 + 50 epochs.

Model	Set	IoU			Accuracy		
		Avg.	Flood	Non-Flood	Avg.	Flood	Non-Flood
U-Prithvi	Test	89.73	82.21	97.24	94.81	91.15	98.46
U-Prithvi (No RHM layer)	Test	89.35	81.54	97.17	94.14	89.69	98.59
U-Prithvi	Bolivia	87.70	79.68	95.71	93.31	88.84	97.78
U-Prithvi (No RHM layer)	Bolivia	87.85	79.97	95.72	93.80	90.02	97.57

4.4 Prithvi Architecture Fine-Tuning

This section addresses our fourth research question and describes several additional experiments conducted to investigate fine-tuning the architectures. We conducted a series of experiments with various parameter configurations to find the optimal performance of our model. The following sections provide a detailed analysis of these parameters and their impact on the model's performance.

4.4.1 Balancing Learning Between Prithvi Features and U-Net Features

The proposed U-Prithvi architecture creates a bottleneck by concatenating the encoded features from both the U-Net and the Prithvi encoder. By merging these two representations, we aim to enable the model to learn effectively from the training data while also generalizing well to out-of-sample data. However, since the Prithvi model is pre-trained while the U-Net component is not, there is a risk that the model may rely solely on one branch of features while ignoring the other, potentially slowing down the learning process.

To mitigate this issue, we introduced the RHM layer, which randomly activates or deactivates the outputs of the Prithvi and U-Net encoders. Specifically, with an equal probability of 1/3, the Prithvi encoder output is masked, the U-Net encoder output is masked, or both remain unchanged. We assess the performance of our U-Prithvi architecture with and without this layer.

Table 3 shows the results of this experiment. Performance with the RHM layer is better when measured using the in-sample test set, but worse when measured using the out-of-sample test set. Based on the results reported on [19], this would be indicative that without the RHM layer, the U-Prithvi model prioritizes the Prithvi encoded features over the U-Net features, since generalization performance is better on in-sample data compared to out-of-sample data. Nonetheless, the difference in performance between the two configurations is much smaller than the difference in performance between the Prithvi-decoder and the U-Prithvi architecture (as shown in Table 1). In addition, the RHM layer makes the model more likely to balance its use of the bottleneck features.

4.4.2 Performance vs. Computational Effort

The U-Prithvi architecture is more complex than both the U-Net and the Prithvi-decoder architecture, which increases the computational effort required to train it. We explore the trade-off between computational effort and predictive performance by changing the number of epochs allowed for training. Table 4 shows the error metrics of the U-Prithvi model when training for different numbers of epochs. We find that a small number of epochs is enough to achieve stable performance in the case of the test set. For instance, the average IoU when

18:12 U-Prithvi: Enhanced Flood Inundation Mapping

■ **Table 4** Performance metrics of the U-Prithvi model for different values of training epochs. All values represent percentages.

Model	Set	Epochs	IoU			Accuracy		
			Avg.	Flood	Non-Flood	Avg.	Flood	Non-Flood
U-Prithvi	Test	25 + 25	89.01	80.87	97.14	92.95	86.97	98.93
U-Prithvi	Test	50 + 50	89.73	82.21	97.24	94.81	91.15	98.46
U-Prithvi	Test	100 + 100	89.71	82.21	97.21	94.03	89.38	98.67
U-Prithvi	Test	200 + 200	89.86	82.34	97.38	93.40	87.74	99.07
U-Prithvi	Bolivia	25 + 25	83.07	71.72	94.42	87.60	76.49	98.72
U-Prithvi	Bolivia	50 + 50	87.70	79.68	95.71	93.31	88.84	97.78
U-Prithvi	Bolivia	100 + 100	87.81	79.68	95.93	93.19	88.28	98.10
U-Prithvi	Bolivia	200 + 200	87.07	78.39	95.75	90.75	82.50	98.99

training 25 + 25 is within 1% of the model trained 200 + 200 epochs. However, a minimum of 50 + 50 is required to achieve most of the model potential in terms of generalization performance.

In both cases, training for more epochs keeps improving the predictive capacity, albeit at much smaller rates. In order to keep our architecture competitive in terms of computational complexity, we will use 50 + 50 epochs of training when comparing it to other models.

4.4.3 Hyper-Parameter Tuning for U-Prithvi

We conducted experiments to determine the optimal hyperparameter configuration for U-Prithvi, testing three additional parameters: the combination operation, the mask probability for the RHM layer, and the ratio of feature embedding sizes between Prithvi and U-Net. Table 5 presents the performance of U-Prithvi under various configurations.

For the combination operation, we evaluated three alternatives: the original strategy of concatenating Prithvi and U-Net embeddings, as well as two additional approaches where embeddings were either summed or multiplied. Notably, these operations are feasible only because the embedding dimensions are identical. Our results indicate that concatenation yields the best performance on the out-of-sample Bolivia set, whereas multiplication performs best on the in-sample test set. However, the performance gains are too small to be statistically significant.

In the second experiment, we explored different probabilities for masking encoder outputs. The default no-mask probability of 33 % means that, with this probability, no part of the encoder outputs was masked. This setting provided the best results for the in-sample test set, whereas a higher no-mask probability of 66% was optimal for the out-of-sample Bolivia set. Similar to the combination operation, the differences between configurations were too minor to be considered statistically significant.

Finally, we examined the effect of varying the embedding size ratio between Prithvi and U-Net. In both test sets, using equal embedding sizes was preferable. We also tested reducing both embeddings by half and doubling each dimension in separate experiments. While performance on the Bolivia set dropped significantly when embedding sizes were halved, it improved only slightly when the Prithvi embedding dimension was doubled.

5 Conclusion

While the advent of geospatial foundational models will undoubtedly give way to novel and innovative methods for analyzing satellite imagery, work is still being done to assess their usefulness against CNN-based models. We contribute to a better understanding of this problem by not only comparing the strengths and weaknesses of the two approaches but also by suggesting a synergistic pathway that can leverage both through our novel U-Prithvi model. Our experiments for flood inundation mapping helped us answer our research questions. First, custom models such as U-Net outperform foundation models like Prithvi on in-sample data, yet the opposite is true for out-of-sample data. Then, we find that combining both architectures improves the performance of the Prithvi model on in-sample data without decreasing its capacity to generalize. Next, increasing the complexity of the U-Net model can produce results matching those of Prithvi. Finally, the proposed RandomHalfMask Layer produces significant performance improvements by balancing the learning between U-Net and Prithvi features. However, careful fine-tuning is necessary for this to yield optimal results. Based on this, we believe that combining a CNN-based deep learning model with a transformer-based foundational model, such as the one we've proposed with U-Prithvi, allows us to leverage the strengths of architectures like U-Net along with the generalizability of a foundation model. U-Net's multi-scale feature learning helps capture

■ **Table 5** Results of hyper-parameter tuning experiments for U-Prithvi.

Bolivia set		IoU			Accuracy		
Parameter		Avg	Floods	Non-Floods	Avg	Floods	Non-Floods
Combination operation	Add	86.97	78.24	95.69	90.82	82.76	98.88
	Multiply	86.39	77.40	95.38	91.27	84.26	98.29
	Concat	87.70	79.68	95.71	93.31	88.84	97.78
No-mask probability	66%	88.14	80.25	96.02	92.27	85.91	98.64
	33%	87.70	79.68	95.71	93.31	88.84	97.78
	16%	87.95	79.96	95.94	92.29	86.07	98.52
Prithvi / U-Net Ratio	0.5 : 0.5	85.51	76.01	95.01	91.06	84.20	97.92
	1 : 2	87.07	78.34	95.80	90.34	81.45	99.23
	1 : 1	87.70	79.68	95.71	93.31	88.84	97.78
	2 : 1	87.42	79.34	95.50	94.06	90.94	97.17

Test set		IoU			Accuracy		
Parameter		Avg	Floods	Non-Floods	Avg	Floods	Non-Floods
Combination operation	Add	89.57	81.82	97.31	93.12	87.16	99.08
	Multiply	89.85	82.35	97.36	93.66	88.34	98.97
	Concat	89.73	82.21	97.24	94.81	91.15	98.46
No-mask probability	66%	89.20	81.19	97.22	92.73	86.38	99.09
	33%	89.73	82.21	97.24	94.81	91.15	98.46
	16%	89.33	81.43	97.22	93.27	87.61	98.93
Prithvi / U-Net Ratio	0.5 : 0.5	89.35	81.52	97.17	94.08	89.54	98.61
	1 : 2	89.08	80.94	97.22	92.18	85.08	99.28
	1 : 1	89.73	82.21	97.24	94.81	91.15	98.46
	2 : 1	89.53	81.82	97.23	94.01	89.31	98.70

finer spatial details in image classification, while the extensive training dataset used for the foundation model enhances the fused model's ability to generalize to previously unseen areas in satellite imagery.

This work not only increases our understanding of transformer-based foundation models compared to more commonly used CNN models in the geospatial field, but also enhances our capabilities for flood inundation mapping. The ability of U-Prithvi to accurately capture flood extents while generalizing to any geographic region has important applications for real-time flood mapping and disaster response management. Incorporating a pre-trained, open-source model like Prithvi, which abstracts away the high compute costs of training on massive amounts of satellite imagery, makes flood mapping more accessible and feasible for governments, companies, and non-profit organizations alike to leverage its capabilities. Moreover, our approach could be applied in near-real-time as satellite images are collected to aid in post-hurricane or post-tsunami disaster response to not only understand the extent of potential damage but also to coordinate response efforts to where it is most needed on the ground. Beyond its significance in disaster response, accurately mapping flood extents has significant value for modeling and understanding Earth's biogeochemical cycles, especially as climate change is anticipated to result in more frequent storm surges and increased sea level rise. Knowing exactly when and where areas are flooded, or inundated, has important implications for Earth's carbon cycles. For example, wetland ecosystems emit methane when flooded, a potent greenhouse gas. Thus, understanding the total extent of flooding is vital to modeling and predicting total methane emissions arising from these ecosystems to develop more robust climate mitigation strategies. Overall, the method we developed here can be applied to a variety of environmental, human safety, and climate-related challenges.

Future work could strengthen this model by increasing the amount of data used for training and fine-tuning, incorporating alternate data sources such as SAR data via a multi-modal approach, or by configuring the framework to be able to ingest higher resolution commercial CubeSat data to achieve results at higher spatial and temporal resolutions. This study focuses specifically on the Prithvi model, but future research could explore integrating it with other geospatial foundation models. Moreover, increased testing on the results of foundation models versus other common machine learning and deep learning models would also be beneficial to the community. While the application studied here is flood inundation mapping, we anticipate this framework could be applied to similar remote sensing and geospatial classification tasks such as land cover mapping, fire detection, or crop monitoring with reasonable accuracy.

References

- 1 November 2024. URL: <https://github.com/cloudtostreet/Sen1Floods11>.
- 2 Akhyar Akhyar, Mohd Asyraf Zulkifley, Jaesung Lee, Taekyung Song, Jaeho Han, Chanhee Cho, Seunghyun Hyun, Youngdoo Son, and Byung-Woo Hong. Deep artificial intelligence applications for natural disaster management systems: A methodological review. *Ecological Indicators*, 163:112067, June 2024. doi:10.1016/j.ecolind.2024.112067.
- 3 Reza Azad, Ehsan Khodapanah Aghdam, Amelie Rauland, Yiwei Jia, Atlas Haddadi Avval, Afshin Bozorgpour, Sanaz Karimijafarbigloo, Joseph Paul Cohen, Ehsan Adeli, and Dorit Merhof. Medical image segmentation review: The success of u-net. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10076–10095, 2024. doi:10.1109/TPAMI.2024.3435571.
- 4 Roberto Bentivoglio, Elvin Isufi, Sebastian Nicolaas Jonkman, and Riccardo Taormina. Deep learning methods for flood mapping: a review of existing applications and future research directions. *Hydrology and Earth System Sciences*, 26(16):4345–4378, August 2022. doi:10.5194/hess-26-4345-2022.

- 5 Amrendra Bhushan, Vikas Chandra Goyal, and Arun Lal Lal Srivastav. Greenhouse gas emissions from inland water bodies and their rejuvenation: a review. *Journal of Water and Climate Change*, page jwc2024561, October 2024. doi:10.2166/wcc.2024.561.
- 6 Derrick Bonafilia, Beth Tellman, Tyler Anderson, and Erica Issenberg. Sen1floods11: a georeferenced dataset to train and test deep learning flood algorithms for sentinel-1. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 835–845, June 2020. doi:10.1109/CVPRW50498.2020.00113.
- 7 Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, April 2018. doi:10.1109/TPAMI.2017.2699184.
- 8 Martin Claverie, Junchang Ju, Jeffrey G. Masek, Jennifer L. Dungan, Eric F. Vermote, Jean-Claude Roger, Sergii V. Skakun, and Christopher Justice. The harmonized landsat and sentinel-2 surface reflectance data set. *Remote Sensing of Environment*, 219:145–161, December 2018. doi:10.1016/j.rse.2018.09.002.
- 9 Jesline Daniel, J. T. Anita Rose, F. Sangeetha Francelin Vinnarasi, and Venkatesan Rajinikanth. Vgg-unet/vgg-segnet supported automatic segmentation of endoplasmic reticulum network in fluorescence microscopy images. *Scanning*, 2022(1):7733860, 2022. doi:10.1155/2022/7733860.
- 10 Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. ISSN: 1063-6919. doi:10.1109/CVPR.2009.5206848.
- 11 Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. URL: <http://www.deeplearningbook.org>.
- 12 Boran Han, Shuai Zhang, Xingjian Shi, and Markus Reichstein. Bridging Remote Sensors with Multisensor Geospatial Foundation Models, April 2024. arXiv:2404.01260. doi:10.48550/arXiv.2404.01260.
- 13 Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger Roth, and Daguang Xu. UNETR: Transformers for 3D Medical Image Segmentation, October 2021. arXiv:2103.10504. doi:10.48550/arXiv.2103.10504.
- 14 Xin He, Yong Zhou, Jiaqi Zhao, Di Zhang, Rui Yao, and Yong Xue. Swin Transformer Embedding UNet for Remote Sensing Image Semantic Segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2022. doi:10.1109/TGRS.2022.3144165.
- 15 Chang Huang, Yun Chen, Shiqiang Zhang, and Jianping Wu. Detecting, extracting, and monitoring surface water from space using optical sensors: A review. *Reviews of Geophysics*, 56(2):333–360, 2018. doi:10.1029/2018RG000598.
- 16 Intergovernmental Panel On Climate Change (Ipcc). *Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, July 2023. doi:10.1017/9781009157896.
- 17 Johannes Jakubik, Sujit Roy, C. E. Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, Daiki Kimura, Naomi Simumba, Linsong Chu, S. Karthik Mukkavilli, Devyani Lambhate, Kamal Das, Ranjini Bangalore, Dario Oliveira, Michal Muszynski, Kumar Ankur, Muthukumaran Ramasubramanian, Iksha Gurung, Sam Khallaghi, Hanxi, Li, Michael Cecil, Maryam Ahmadi, Fatemeh Kordi, Hamed Alemohammad, Manil Maskey, Raghu Ganti, Kommy Weldemariam, and Rahul Ramachandran. Foundation models for generalist geospatial artificial intelligence. *CoRR*, November 2023. arXiv:2310.18660. doi:10.48550/arXiv.2310.18660.
- 18 A. Emily Jenifer and Sudha Natarajan. Deepflood: A deep learning based flood detection framework using feature-level fusion of multi-sensor remote sensing images. *JUCS - Journal of Universal Computer Science*, 28(33):329–343, March 2022. doi:10.3897/jucs.80734.


- 19 Wenwen Li, Hyunho Lee, Sizhe Wang, Chia-Yu Hsu, and Samantha T. Arundel. Assessment of a new geoi foundation model for flood inundation mapping. *CoRR*, November 2023. arXiv:2309.14500. doi:10.48550/arXiv.2309.14500.
- 20 Zhouyayan Li and Ibrahim Demir. U-net-based semantic classification for flood extent extraction using sar imagery and gee platform: A case study for 2019 central us flooding. *Science of The Total Environment*, 869:161757, April 2023. doi:10.1016/j.scitotenv.2023.161757.
- 21 Yuchi Ma, Shuo Chen, Stefano Ermon, and David B. Lobell. Transfer learning in environmental remote sensing. *Remote Sensing of Environment*, 301:113924, February 2024. doi:10.1016/j.rse.2023.113924.
- 22 Gonzalo Mateo-Garcia, Joshua Veitch-Michaelis, Lewis Smith, Silviu Vlad Oprea, Guy Schumann, Yarin Gal, Atılım Güneş Baydin, and Dietmar Backes. Towards global flood mapping onboard low cost satellites with machine learning. *Scientific Reports*, 11(1):7249, March 2021. doi:10.1038/s41598-021-86650-z.
- 23 Matias Mendieta, Boran Han, Xingjian Shi, Yi Zhu, and Chen Chen. Towards Geospatial Foundation Models via Continual Pretraining, August 2023. arXiv:2302.04476. doi:10.48550/arXiv.2302.04476.
- 24 Yohei Nakayama, Jiawei Su, and Luis M. Pazos-Outón. SatSwinMAE: Efficient Autoencoding for Multiscale Time-series Satellite Imagery, October 2024. arXiv:2405.02512. doi:10.48550/arXiv.2405.02512.
- 25 Maxim Neumann, Andre Susano Pinto, Xiaohua Zhai, and Neil Houlsby. In-domain representation learning for remote sensing, November 2019. arXiv:1911.06721. doi:10.48550/arXiv.1911.06721.
- 26 Keiller Nogueira, Otávio A. B. Penatti, and Jefersson A. dos Santos. Towards better exploiting convolutional neural networks for remote sensing scene classification. *Pattern Recognition*, 61:539–556, January 2017. doi:10.1016/j.patcog.2016.07.001.
- 27 Fernando Pech-May, Raúl Aquino-Santos, Omar Álvarez Cárdenas, Jorge Lozoya Arandia, and German Rios-Toledo. Segmentation and visualization of flooded areas through sentinel-1 images and u-net. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:8996–9008, 2024. doi:10.1109/JSTARS.2024.3387452.
- 28 Olivier Petit, Nicolas Thome, Clément Rambour, and Luc Soler. U-Net Transformer: Self and Cross Attention for Medical Image Segmentation, March 2021. arXiv:2103.06104. doi:10.48550/arXiv.2103.06104.
- 29 Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, May 2015. arXiv:1505.04597 [cs]. arXiv:1505.04597.
- 30 Rizwan Sadiq, Zainab Akhtar, Muhammad Imran, and Ferda Ofli. Integrating remote sensing and social sensing for flood mapping. *Remote Sensing Applications: Society and Environment*, 25:100697, January 2022. doi:10.1016/j.rsase.2022.100697.
- 31 Apoorva Shastry, Elizabeth Carter, Brian Coltin, Rachel Sleeter, Scott McMichael, and Jack Eggleston. Mapping floods from remote sensing data and quantifying the effects of surface obstruction by clouds and vegetation. *Remote Sensing of Environment*, 291:113556, June 2023. doi:10.1016/j.rse.2023.113556.
- 32 Nahian Siddique, Sidike Paheding, Colin P. Elkin, and Vijay Devabhaktuni. U-net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access*, 9:82031–82057, 2021. doi:10.1109/ACCESS.2021.3086020.
- 33 Zhongbin Su, Wei Li, Zheng Ma, and Rui Gao. An improved u-net method for the semantic segmentation of remote sensing images. *Applied Intelligence*, 52(3):3276–3288, February 2022. doi:10.1007/s10489-021-02542-9.
- 34 Mirela G. Tulbure, Mark Broich, Vinicius Perin, Mollie Gaines, Junchang Ju, Stephen V. Stehman, Tamlin Pavelsky, Jeffrey G. Masek, Simon Yin, Joachim Mai, and Luc Betbeder-Matibet. Can we detect more ephemeral floods with higher density harmonized landsat sentinel 2 data compared to landsat 8 alone? *ISPRS Journal of Photogrammetry and Remote Sensing*, 185:232–246, March 2022. doi:10.1016/j.isprsjprs.2022.01.021.

- 35 Yuefei Wang, Xi Yu, Yixi Yang, Shijie Zeng, Yuquan Xu, and Ronghui Feng. FTUNet: A Feature-Enhanced Network for Medical Image Segmentation Based on the Combination of U-Shaped Network and Vision Transformer. *Neural Processing Letters*, 56(2):83, March 2024. doi:10.1007/s11063-024-11533-z.
- 36 Michael A. Wulder, Jeffrey G. Masek, Warren B. Cohen, Thomas R. Loveland, and Curtis E. Woodcock. Opening the archive: How free data has enabled the science and monitoring promise of landsat. *Remote Sensing of Environment*, 122:2–10, July 2012. doi:10.1016/j.rse.2012.01.010.
- 37 Qiming Yang, Zixin Wang, Shinan Liu, and Zizheng Li. Research on improved u-net based remote sensing image segmentation algorithm. In *2024 6th International Conference on Internet of Things, Automation and Artificial Intelligence (IoTAAI)*, pages 686–689, July 2024. doi:10.1109/IoTAAI62601.2024.10692547.
- 38 Bofei Zhao, Haigang Sui, and Junyi Liu. Siam-dwenet: Flood inundation detection for sar imagery using a cross-task transfer siamese network. *International Journal of Applied Earth Observation and Geoinformation*, 116:103132, February 2023. doi:10.1016/j.jag.2022.103132.


Search Space Reduction Using Species Distribution Modeling with Simulated Pollen Signatures

Haoyu Wang 

Department of Geography and the Environment, University of Texas at Austin, TX, USA

Jennifer A. Miller 

Department of Geography and the Environment, University of Texas at Austin, TX, USA

Shalene Jha 

Department of Integrative Biology, University of Texas at Austin, TX, USA

Abstract

Microscopic trace materials, such as pollen, are an important category of forensic evidence recovered during investigations. As an environmentally ubiquitous substance that can attach to various surfaces, pollen enables the linking of objects and people in space and time. In this study, we assessed the extent to which the search space could be reduced using simulated pollen signatures. These signatures were compiled by randomly selecting pairs of geographic coordinates on the Earth's terrestrial land and querying the Global Biodiversity Information Facility (GBIF) database to identify plant taxa within 50 meters of the coordinates. These taxa were then treated as the parent taxa of the pollen, simulating the hypothetical attachment of pollen signatures to objects or individuals. For each identified pollen taxon, we modeled habitat suitability for the parent plant taxa and combined the spatial distributions to refine the geolocation search area. Since the actual coordinates for these locations of interest were known, we were able to evaluate the global performance of the search space reduction under the assumption of an extreme constraint that no other contextual information was available.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases geoforensics, species distribution modeling, search space reduction

Digital Object Identifier 10.4230/LIPICs.GIScience.2025.19

Funding The research was partly funded by DEVCOM ARL, ARO through a Multidisciplinary University Research Initiative Grant (#W911NF1910231). The research, interpretations, and perspectives reported here are those of the authors and should not be attributed to the Army or the Department of Defense.

1 Introduction

Geoforensic applications incorporate data collection and analytical methods from spatial data science, remote sensing, and earth sciences to aid forensic investigations in environmental issues, criminal justice, and human rights [5]. Pollen grains and signatures are suitable candidates for trace evidence retrieved during investigations in geoforensic analyses because of their environmental ubiquity and durability [6, 4, 2]. DNA metabarcoding with high-throughput sequencing technologies dramatically improved pollen identification in quantity and taxonomic accuracy, leading to potentially more reliable applications of such forensic evidence (Bell et al. 2016). New research using environmental DNA samples such as pollen in species distribution models (SDM) for geoforensic location analysis has shown promise [8]. These models are used to quantify species-environment correlations which can then be used to predict the habitat suitability or potential plant species distribution in a geographic information system [3].



© Haoyu Wang, Jennifer A. Miller, and Shalene Jha;
licensed under Creative Commons License CC-BY 4.0

13th International Conference on Geographic Information Science (GIScience 2025).

Editors: Katarzyna Sila-Nowicka, Antoni Moore, David O'Sullivan, Benjamin Adams, and Mark Gahegan;
Article No. 19; pp. 19:1–19:6



Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

In this study, we use an SDM framework to reduce the search space of geographic location(s) associated with an object (e.g., laptop, clothes, person) and use simulated pollen signatures to test its applicability. The simulated pollen signatures were obtained by generating random coordinates as locations of interest worldwide and then downloading plant species data from the Global Biodiversity Information Facility (GBIF) within a specified distance from these locations to establish simulated pollen signatures. We then considered these locations of interest as the locations that an object traveled through. For each pollen signature, we estimated SDMs for its parent taxa and used a scaled-sum method to combine the relative suitability of the different plant species associated with the object. Since the geographic coordinates for these locations of interest were known for simulated examples, we could evaluate the performance of the search space reduction (or search score) under the assumption of an extreme constraint that only the pollen taxa, serving as the trace material, is known. Investigators typically have a general understanding of the potential activity boundaries of individuals in a given case. However, this information may not always be available, especially for international cases where people travel and objects move between continents. Thus, this research usually assumes two scenarios for the objects of interest:

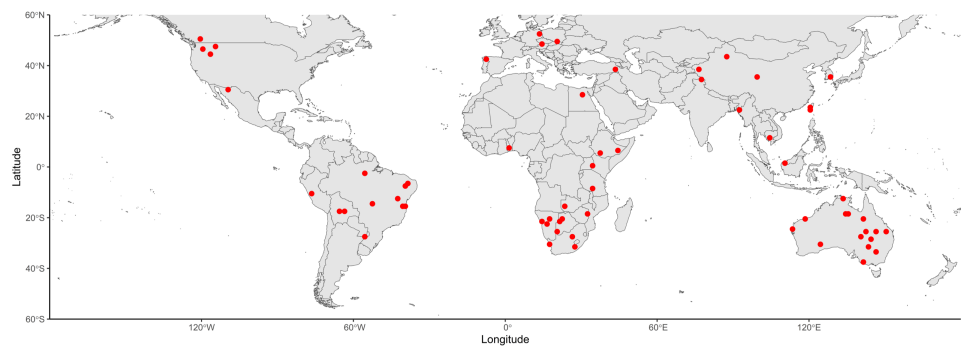
1. We have limited information about the objects of interest; for example, they are traveling within the United States
2. There is no information on any location history of the objects of interest *a priori*

For the second scenario, investigators may need to set larger potential areas of interest. The modeling approach requires processing a large amount of data during geographic attribution at a large spatial extent. As a result, such efforts pose a data challenge that is more computationally demanding. To deal with this challenge, we build the geographic attribution workflow at a global extent using Google Earth Engine (GEE), a cloud-based geospatial platform designed to support large-scale modeling across broad spatial extents with petabyte-scale data access and fast computation.

2 Methods

We randomly generated 9,999 points on global terrestrial land and queried GBIF for plant occurrences within 50 meters of each point. These queries generated 65 locations of interest that have at least more than one plant taxa found within the 50-meter distance threshold. We here assumed a 100% probability of pollen adhering to objects or individuals. In other words, we assumed objects traveled to this location and picked up pollen grains from their parent plants. There are 246 unique plant taxa associated with the 65 locations of interest. The data structure of the simulated pollen signatures can be found in Table 1. While the locations and plant species used in this approach are real, we use the word simulation to refer to the process of collecting the object and identifying plant species from it in comparison to one based on fieldwork sampling.

To strengthen the comparison between simulated and real-world conditions, we incorporated sampled pollen signatures as a reference, we also included the previously calculated search space reduction scores from fieldwork-sampled pollen signatures. The sampled pollen signatures were directly collected using sampling instruments in the great Austin area, Texas, with different sampling methods including air pollen samplers and stationary fabric samples. Pollen grains were then identified in the laboratory with light microscopy and DNA metabarcoding. This process then approximates the real-world attaching of pollen onto different surfaces. With sampled pollen signatures, we can further compare search space reduction results by incorporating both simulated and real-world sampling methods.



■ **Figure 1** The 65 simulated sites out of 9,999 sites that has plant taxa recorded in GBIF within a 50 m distance. The base map shows the near-global extent used for modeling in this study. At a 900 m spatial resolution, there are around 176.22 million cells in this near-global study area.

■ **Table 1** Example data structure on the information of simulated pollen attachment to randomly generated locations of interest.

Simulated Pollen Signatures				
Locations of Interest	<i>L. ramosissima</i>	<i>M. citriodora</i>	<i>A. columbianum</i>	...
(-114.552, 47.547)	1	1	1	...
(92.588, 22.596)	0	0	1	...
(-55.542, -27.573)	0	1	0	...
...			...	

We also used a null model approach as a baseline to compare how well our geographic attribution method performed against random chance. In the null model, we assigned the same number of pollen taxa to randomly chosen locations, but instead of selecting them based on geographic proximity (within 50 meters of a known location), we randomly picked them from the overall pool of simulated pollen signatures. This approach helps determine whether the geographic patterns we observed in the simulated data are meaningful or if they could have occurred by random chance. By comparing the results of the simulated pollen signatures to those from the null model, we can evaluate whether our method provides useful search space reduction beyond what would be expected randomly.

The spatial extent of the SDMs in this study was terrestrial land on all continents except Antarctica, from -180° W to 180° E and -60° S to 60° N as shown in Figure 1. The SDM-based geographic attribution workflow is open-source and implemented using the Google Earth Engine (GEE) Python API. We selected only georeferenced occurrences with locations on terrestrial land with an occurrence limit of 5,000 using a programmatic interface *rgbif* that queries species occurrence records [1]. Since the occurrence data from open-source global databases such as GBIF does not have associated absence data, the SDM requires other forms of absence information, such as samples of background or pseudo-absence data. Considering the computational feasibility of fitting SDMs on GEE, we set a spatial resolution of 900 m for this study and created pseudo-absences on cells that are less similar in terms of the environmental conditions to the cells with presence data using the *k*-means clustering method. We fit SDMs with two methods, Random Forest (RF) and Boosted Regression Trees (BRT), using the machine learning classifiers in the Statistical Machine Intelligence and Learning Engine (SMILE) available on GEE. The training and testing processes were implemented on 500×500 km spatial blocks with an 80/20 training/testing data split. The Area under

the Receiver Operating Characteristic Curve (AUC_{ROC}) was used as the SDM classification performance metric. Variables related to temperature, precipitation, and elevation were selected as environmental predictors for modeling habitat suitability in this study. To get temperature and precipitation variables, 19 bioclimatic layers at a 30 arc-second spatial resolution from the WorldClim V1 Bioclim dataset were retrieved in GEE (Hijmans et al. 2005).

After fitting SDMs, the next step is to combine the distribution maps to estimate the geolocation of the object when multiple species have been identified. Studies have used joint probabilistic approaches for target distribution estimation, often assuming independent occurrences of taxa, but a zero probability from any taxa can incorrectly exclude locations unless mitigated by techniques like setting a minuscule probability. To address this problem, we followed [7] and [8] and used a scaled-sum method to generate joint suitability maps to combine single-taxa SDM prediction results of the objects with more than one pollen taxa recovered that maintain relative suitabilities. To achieve this, we use notations i, j as indices of longitude/latitude pixels on a species suitability map with $M \times N$ total pixels, where $i = 1, \dots, M$ and $j = 1, \dots, N$. We also employ k as the genus/species of pollen identified on the target object, where $k = 1, \dots, n$, if n types of pollen taxa are recovered. For an SDM-generated suitability distribution of a plant taxon, we can generate a suitability matrix \mathbf{P}^k . The $p_{i,j}$ is a single-taxa suitability score generated from SDMs at a location (i, j) in the study extent. We can then derive a joint suitability distribution map \mathbf{S} for each object of interest. We define a percentile approach in environmental space to evaluate the geographic attribution results of the objects. We can link the joint suitability maps by computing the percentile of numeric pixel values from all cells for each map. Since each joint suitability map of geographic attribution for every modeling method is only dependent on the suitability distributions \mathbf{P}^k in a given study area extent, whether a modeling method can or cannot identify an object at a numeric percentile is an evaluation metric comparable across methods and sampling regions. If we use \mathbf{P}^k layers to generate a joint suitability map \mathbf{S} , then for each joint suitability layer of an object of interest, we have the percentile for each pixel value on \mathbf{S} , ranging from $[0, 100]$. A pixel with a higher percentile indicates a greater likelihood that the object of interest has traveled in or around this location, analogous to suitability in spatial modeling. Specifically, we can calculate a percentile of the sampling location of an object S_0 to assess the geolocation accuracy resulting from the geographic attribution. We call this specific value a search space reduction (SSR) score:

$$SSR = \left(1 - \frac{\#\{(i, j) : [\frac{1}{n} \sum_{k=1}^n \mathbf{P}^k]_{ij} \geq S_0\}}{M \cdot N} \right) \cdot 100 \quad (1)$$

SSR score is a metric that can retrospectively assess how well the method could reduce the search space by comparing the joint suitability value between the object's location and all other locations. Equation 1 produces SSR scores between 0 and 100, where a higher score indicates that fewer pixels on the joint suitability map have suitability scores greater than or equal to that of the object's actual location.. We use this concept in this research to evaluate the performance of the geographic attribution process. It is important to stress here again that this SSR score based upon the joint suitability method is a proxy for location suitability to identify location history.

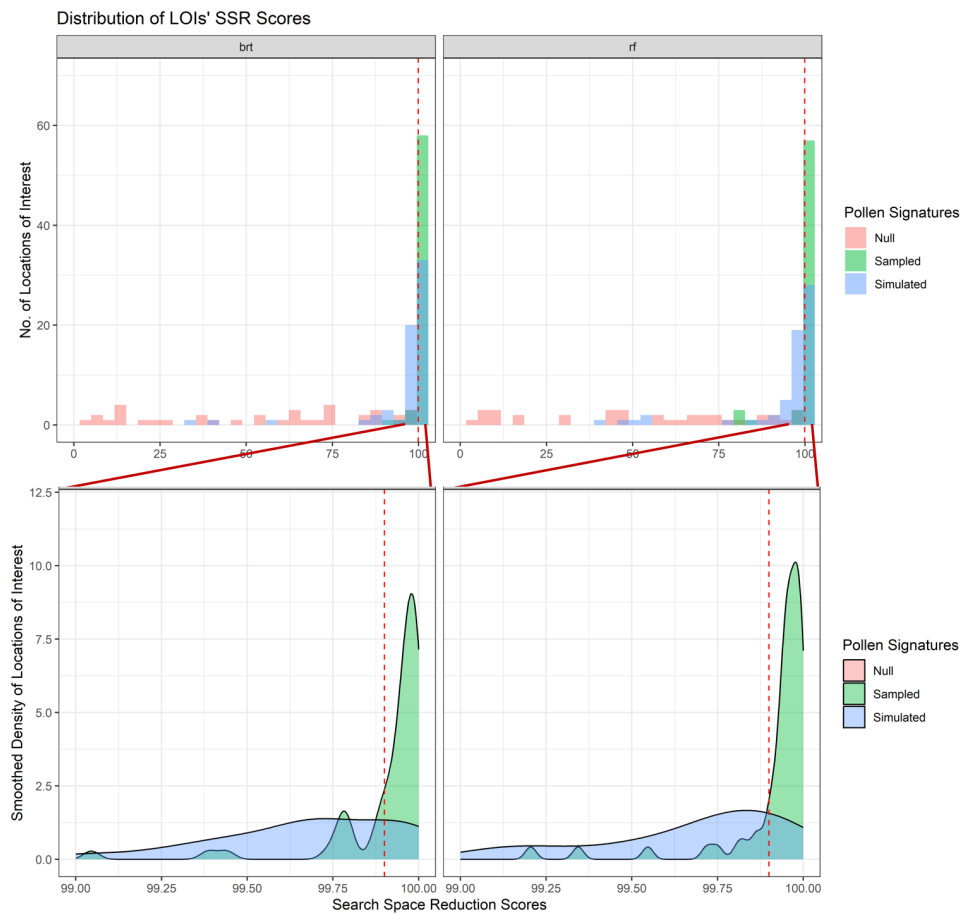


Figure 2 The distribution of SSR scores resulted from using sampled, simulated, and null model pollen signatures. Both sampled and simulated methods have 65 locations of interest selected. We used 38 locations of interest for the pollen signatures generation of the null model.

3 Results and Discussion

AUC_{ROC} results of 0.75 or higher for 90% of the plant taxa classifications indicate overall useful predictive performance from SDMs. For the simulated sites, both BRT and RF models resulted in a search space reduction (SSR) score > 99.95 for around half of the locations. Figure 2 shows the distribution of the search scores of the locations of interest derived from simulated pollen signatures and sampled pollen signatures shown in blue and green color. The null model distribution with randomly assigned pollen signatures is in red. The distribution of most null model SSR scores is below the 75 search space percentile, which means they do not provide useful information for search space reduction. The distribution of the SSR scores yielded from simulated pollen signatures is concentrated at the 99th percentile, while the majority of the locations of interest have the highest SSR scores with sampled pollen signatures, which can also be noted in the zoomed-in smoothed density plot in Figure 2.

To provide a more detailed analysis, we used an additional sub-figure that highlights the upper-end distribution of SSR scores using both BRT and RF models. The zoomed-in density plots in the lower panel of Figure 2 emphasize the peak concentration of sampled and simulated pollen signatures at the extreme high SSR scores. This supports the hypothesis

that geolocation using sampled and simulated pollen provides significantly better search space reduction than random pollen attachment, as represented by the null model. The BRT model demonstrates a slightly broader spread of high SSR scores compared to the RF model, potentially indicating model-specific differences in how species distribution models generalize habitat suitability.

SSR scores derived from simulated pollen signatures can be compared with the geographic attribution results from sampled pollen signatures. To introduce uncertainty in pollen adherence, we can adjust the probability of pollen attachment to objects/people to values less than 1, allowing us to assess the sensitivity of the SSR modeling to that parameter. Although multiple potential search regions can be identified, investigators and decision-makers could use these refined maps to reference location history at higher percentiles of areas of interest, especially when combined with other lines of evidence. Future analyses should explore the robustness of these distribution patterns across different geographic extents, alternative modeling techniques, and additional environmental variables to assess their impact on SSR performance. Additionally, integrating higher-resolution pollen data or refining taxonomic resolution may further enhance the precision of location attribution in forensic geospatial investigations.

References

- 1 Scott Chamberlain, Vijay Barve, Dan McGlinn, Damiano Oldoni, Peter Desmet, Laurens Geffert, and Karthik Ram. Rgbif: Interface to the Global Biodiversity Information Facility API, April 2022.
- 2 Edward Helderop, Tony H. Grubestic, Elisa Jayne Bienenstock, Skaidra Smith-Heisters, Haoyu Wang, and Jennifer A. Miller. Geoforensic Palynology Search Models and Human-Mediated Secondary Pollen Deposition. *The Professional Geographer*, 77(2):1–13, 2025. doi:10.1080/00330124.2024.2434473.
- 3 Jennifer A. Miller. Species distribution models: Spatial autocorrelation and non-stationarity. *Progress in Physical Geography: Earth and Environment*, 36(5):681–692, October 2012. Publisher: SAGE Publications Ltd. doi:10.1177/0309133312442522.
- 4 Wangshu Mu, Daoqin Tong, Tony H. Grubestic, Hung-Chi Liu, Edward Helderop, Jennifer A. Miller, and Elisa Jayne Bienenstock. Geoforensics with Pollen Quantification: A Spatial Perspective. *Annals of the American Association of Geographers*, 113(9):1–17, 2023. doi:10.1080/24694452.2023.2211155.
- 5 Alastair Ruffell and Jennifer McKinley. *Geoforensics*. John Wiley & Sons, October 2008. Google-Books-ID: f3UCEAAAQBAJ.
- 6 Libby A. Stern, Jodi B. Webb, Debra A. Willard, Christopher E. Bernhardt, David A. Korejwo, Maureen C. Bottrell, Garrett B. McMahon, Nancy J. McMillan, Jared M. Schuetter, and Jack Hietpas. Geographic Attribution of Soils Using Probabilistic Modeling of GIS Data for Forensic Search Efforts. *Geochemistry, Geophysics, Geosystems*, 20(2):913–932, 2019. doi:10.1029/2018GC007872.
- 7 Haoyu Wang, Jennifer A. Miller, Tony H. Grubestic, and Shalene Jha. A Framework for Using Ensemble Species Distribution Models for Geographic Attribution in Forensic Palynology. In *2022 IEEE International Symposium on Technologies for Homeland Security (HST)*, pages 1–7, November 2022. doi:10.1109/HST56032.2022.10025427.
- 8 Haoyu Wang, Jennifer A. Miller, Tony H. Grubestic, and Shalene Jha. Using habitat suitability models for multiscale forensic geolocation analysis. *Transactions in GIS*, 27(3):777–796, 2023. doi:10.1111/tgis.13052.