

Generalized spatial association modeling driven by the nature of geospatial data

Peng Luo

Vollständiger Abdruck der von der TUM School of Engineering and Design der
Technischen Universität München zur Erlangung eines

Doktors der Ingenieurwissenschaften (Dr.-Ing.)

genehmigten Dissertation.

Vorsitz:

Univ.-Prof. Dr. Walter Timo de Vries

Prüfende der Dissertation:

1. Prof. Dr.-Ing. Liqiu Meng
2. Prof. Dr. Franz-Benjamin Mocnik
3. Prof. Yu Liu

Die Dissertation wurde am 25.03.2024 bei der Technischen Universität München
eingereicht und durch die TUM School of Engineering and Design am 17.06.2024
angenommen.

Abstract

Analyzing geospatial data is crucial for various domains such as urban planning, environmental management, and public health. Geospatial data, as a unique form of data, recording the spatial attributes and interactions of different locations, making its accurate and comprehensive description pivotal for the success of spatial data analysis. The success in representing and analyzing geospatial information is key to ensuring the accuracy and trustworthiness of findings in these areas, calling for a careful and detailed approach to describing spatial data¹. This research emphasizes the importance of a refined understanding and modeling of spatial data, addressing the complexities and specificities that distinguish it from other data types, to enhance the precision and effectiveness of spatial analysis.

The objective of this thesis is to develop a more comprehensive understanding of spatial data's uniqueness and to leverage this understanding to improve spatial analysis methods. This thesis focuses on the challenges in spatial data analysis across three distinct tasks (i.e. descriptive tasks, explanatory tasks, and predictive tasks), stemming from the incomplete description of spatial data characteristics, and endeavors to address these issues through targeted investigations.

The thesis is structured around three main attempts to address these challenges:

Identifying overlapping community structures: It explores methods to detect overlapping community structures within spatial interaction data, recognizing that these structures can significantly influence spatial dynamics and interactions.

Explaining nonlinear interactions between geographical variables: The research investigates the complex, nonlinear relationships between geographical variables, aiming to develop models that can better account for these dynamics under weak statistical assumptions.

Improving spatial prediction in sparse and biased samples: The dissertation focuses on refining spatial prediction methods to overcome the limitations posed by sparse and biased samples, enhancing the reliability of predictions in spatial analysis.

Through these efforts, the dissertation seeks to bridge the research gap by providing a more accurate and nuanced description of spatial data, thereby facilitating more effective spatial analysis and decision-making processes.

In this thesis, "spatial" will be used in the context as a synonym for "geospatial".

Zusammenfassung

Die Analyse von Geodaten ist für verschiedene Bereiche wie Stadtplanung, Umweltmanagement und öffentliche Gesundheit von entscheidender Bedeutung. Geodaten, als eine einzigartige Form von Daten, die die räumlichen Attribute und Interaktionen verschiedener Orte aufzeichnet, machen ihre genaue und umfassende Beschreibung entscheidend für den Erfolg der räumlichen Datenanalyse. Der Erfolg bei der Darstellung und Analyse von geografischen Informationen ist entscheidend, um die Genauigkeit und Vertrauenswürdigkeit von Ergebnissen in diesen Bereichen sicherzustellen, was eine sorgfältige und detaillierte Herangehensweise an die Beschreibung von räumlichen Daten erfordert. Diese Forschung betont die Bedeutung eines verfeinerten Verständnisses und Modellierung von räumlichen Daten, um die Komplexitäten und Spezifika zu adressieren, die es von anderen Datentypen unterscheiden, und um die Präzision und Effektivität der räumlichen Analyse zu verbessern.

Das Ziel dieser Arbeit ist es, ein umfassenderes Verständnis der Einzigartigkeit räumlicher Daten zu entwickeln und dieses Verständnis zu nutzen, um räumliche Analysemethoden zu verbessern. Diese Dissertation konzentriert sich auf die Herausforderungen in der räumlichen Datenanalyse in drei verschiedenen Aufgaben (d.h. beschreibende Aufgaben, erklärende Aufgaben und prädiktive Aufgaben), die aus der unvollständigen Beschreibung der räumlichen Datenmerkmale resultieren, und bemüht sich, diese Probleme durch gezielte Untersuchungen anzugehen.

Die Arbeit ist um drei Hauptansätze strukturiert, um diese Herausforderungen anzugehen:

Identifizierung sich überlappender Gemeinschaftsstrukturen: Es werden Methoden untersucht, um sich überlappende Gemeinschaftsstrukturen innerhalb von räumlichen Interaktionsdaten zu erkennen, wobei erkannt wird, dass diese Strukturen die räumlichen Dynamiken und Interaktionen wesentlich beeinflussen können.

Erklärung nichtlinearer Wechselwirkungen zwischen geografischen Variablen: Die Forschung untersucht die komplexen, nichtlinearen Beziehungen zwischen geografischen Variablen, mit dem Ziel, Modelle zu entwickeln, die diese Dynamiken unter schwachen statistischen Annahmen besser berücksichtigen können.

Verbesserung der räumlichen Vorhersage in spärlichen und voreingenommenen Stichproben: Die Dissertation konzentriert sich darauf, räumliche Vorhersagemethoden zu verfeinern, um die durch spärliche und voreingenommene Stichproben verursachten Einschränkungen zu überwinden und die Zuverlässigkeit von Vorhersagen in der räumlichen Analyse zu verbessern.

Durch diese Bemühungen zielt die Dissertation darauf ab, die Forschungslücke zu überbrücken, indem sie eine genauere und nuanciertere Beschreibung räumlicher Daten bereitstellt, was eine effektivere räumliche Analyse und Entscheidungsfindungsprozesse ermöglicht.

Acknowledgment

故乡遥，何日去。
家住吴门，久作长安旅。
五月渔郎相忆否。
小楫轻舟，梦入芙蓉浦。

Whenever a child from my relatives' family, still in elementary or middle school, didn't get a good grade, my mother would always comfort them by saying, "Don't worry, even someone as not-so-smart as Peng managed to get into college." I tried to argue, but my mother would just smile and ignore me. Gradually, I realized that, since she knew me better than I knew myself, I had no choice but to accept it calmly. Now that I've earned a PhD, I feel more qualified to encourage younger people who are still struggling in their studies. Although earning a PhD is undoubtedly challenging, don't worry—if someone as not-so-smart as me could manage it, so can you.

On February 12, 2024, the day after I started writing my doctoral dissertation, my mother left me. I can hardly remember how I managed to complete the dissertation during that time. Throughout those years of pursuing my PhD, I was almost entirely consumed by my research. In that dark moment, when life became unbearable, research became my last refuge—an escape from overwhelming sorrow. After my PhD defense, I found it nearly impossible to write this acknowledgment section. It took me two months to complete my dissertation but six months to gather the courage to write these words.

My initial plan was to dedicate this acknowledgment entirely to my mother, but I realized that it was an impossible task. An acknowledgment reflects a person's experiences during a particular period of life and the people they encountered along the way, but a person's love for their mother can never be summarized. In addition, I found myself almost lacking the courage to face a world without her. There were moments when I wished to bid farewell to everyone—and to this world—without saying goodbye. But I came to realize that I couldn't, because I had already accumulated so many debts of gratitude.

Nonetheless, I must begin by thanking my mother, Mrs. Xiumei Guo. I was born into a poor rural family in China, where both of my parents were migrant workers. They only completed elementary or middle school before entering the workforce. Despite having little, they provided everything I needed to grow, both materially and spiritually. To raise me, my mother essentially became a full-time caregiver, which made our family's financial situation even more difficult. During my childhood, she tried her best to supplement our income whenever she had a moment outside of caring for me. Although my memories are faint, I still recall her taking me to sell breakfast outside my elementary school, trimming excess threads from clothes at the factory, delivering goods to the hospital, and working as a construction laborer. I will never forget the time I video-called her from Germany and saw her wearing paint-stained clothes, sitting in a room mid-renovation. In that moment, I knew I had no reason to stop working—not out of any noble or lofty ideal, but simply from the hope that one day my mother and family would no longer have to struggle.

But fate never gave me the chance to take care of my mother. I neither had the right nor the fortune to give her a good life. She brought me into this world and raised me, but she could only stay by my side for 28 years. What saddens me even more is that, in those 28 years, I never had the chance to say thank you. So, first and foremost, I thank my mother for giving me life. She spent 28 years loving me, and I will spend the rest of my life drawing strength from that love. Though you are gone, I feel you living within me, in every moment of my life.

I would like to express my deepest gratitude to my PhD advisor, Prof. Liqiu Meng, for providing me with the best support and guidance an advisor could offer. During my three and a half years as a PhD student, I often took pride in my small accomplishments. But over time, I came to realize that these achievements were only possible because I was fortunate enough to receive her unwavering support. Prof. Meng always prioritized her students' needs and encouraged me to pursue any direction that interested me. She upheld exceptionally high academic standards, though she never imposed them on me. Instead, she gave me the freedom to explore while offering subtle yet essential guidance. If, at the end of my PhD journey, I have developed a decent academic taste, it is entirely due to her influence.

I am also deeply grateful to Prof. Mocnik from the University of Salzburg and Prof. Liu Yu from Peking University, my committee members, for their invaluable advice. I will always cherish the day Prof. Mocnik warmly welcomed me in Salzburg. We spent an afternoon immersed in research discussions—one of the most enjoyable academic conversations I had during my PhD journey. Prof. Liu was the one who introduced me to the field of GIScience. Four years ago, when I was still a master's student in remote sensing, I attended several of his lectures, where I first learned about the concept of social sensing. Many of my closest friends come from his research group. We studied, played, and drank together, and gradually, I became captivated by GIS—that's how my journey in this field began. I would also like to thank Prof. de Vries for serving as the chair of my dissertation and for his unwavering support. From the moment I submitted my dissertation to my defense, the entire process took only three months. This was far from the usual timeline in graduate school, and it wasn't because my dissertation was exceptional—it was all thanks to Prof. de Vries's help.

During my PhD research, I was fortunate to meet several talented young researchers who provided me with tremendous support. My research focuses on developing better numerical models to help people understand geographic data and the world it represents, ultimately aiming to improve human life on Earth. The direction, values, and philosophy behind my research did not arise solely from personal reflection but were shaped through insightful discussions with these remarkable young scientists. These conversations deeply resonated with me, gradually converging and evolving into research topics that felt almost tailor-made for me. If my research area holds any uniqueness or has made me an expert in a small field, it is purely a matter of luck.

I would like to thank Prof. Yongze Song from Curtin University, who ignited my passion for designing more accurate spatial analysis algorithms. Over the past few years, we both became obsessed with this topic, spending countless hours discussing it. Whenever one of us had an idea, we would call each other immediately, and the discussion would not end until we determined whether the idea was feasible or not. Such discussions made me fervently passionate about academia. I would also like to thank Prof. Di Zhu from the University of Minnesota, Twin Cities. He was the one who brought spatial thinking into my research. His insatiable quest for geographic knowledge inspired me, making me realize that the ultimate goal of geographic research is not to develop the most accurate algorithms, but to develop methods that can uncover geographic knowledge and help people understand the world. I would also like to thank Prof. Ziqi Li from Florida State University. He has a deep understanding and inheritance of traditional spatial analysis methods while also embracing modern AI models. He is committed to building a bridge between tradition and modernity in geographic modeling. His research philosophy inspired my love, acceptance, and tolerance of both the classical and contemporary in geography. Lastly, I want to thank Prof. Yao Yao from China University of Geosciences. He is like my mentor, as well as a close friend. He taught me how to plan my career, how to turn a research idea into a project to be managed, and how to execute every step of the project: from building a research framework, managing data and experiments, writing papers, to visualizing the results. At the same time, he gave me many opportunities to guide students, helping me gradually become an independent researcher.

I would like to express my gratitude to my master's advisor, Prof. Xianfeng Zhang, for his unwavering support and help. Many thanks to the mentors who offered selfless help and advice during my academic

journey: Prof. Song Gao, Dr. Weiming Huang, Prof. Jinfeng Wang, Prof. Fan Zhang, Dr. Zhewei Liu, Prof. Linfang Ding, Prof. Shifen Cheng, Prof. Lei Dong, Prof. Siqin Wang, Prof. Jinfeng Wang, Prof. Fenzhen Su, among others. I have no way to repay your kindness.

I would also like to thank my colleagues and friends in Munich. Thank you, Yu Feng, Zhaiyu Chen, Zhe Zeng, Yan Xia, Dr. Lin Zhou, Chenyu Fang, Yishui Zhou, Bo Zhang. Each gathering with you has added color to my memories of Munich. Thanks to my colleagues at LfK: Nianhua Liu, Chuan Chen, Puzhen Zhang, Mengyi Wei, Xiayin Lou, Shengkai Wang, Dongsheng Chen, Zihan Liu, Hongyi Luo, Xuyang Chen, Jiaying Xue, Christian, Holger, Andreas, Edyta, Bing Liu, Chenyu Zuo, Ruoxin Zhu, Lu Liu, among others. I also thank the best secretary, Stephanie.

I am grateful to all my friends from Oxford. Prof. Jim Hall, Xiao Li, Si Qiao, Alberto Fernández, Deng Majok Chol, Nicholas Chow, Raghav Pant, Tim Fowler, Tom Russell, among others. I will forever cherish that summer spent in Oxford.

Thank you to all my friends in China. Thank you to my friends at Peking University: Jianfeng Huang, Junyi Cheng, Siyuan Liu, Qi Yin, Nengjing Guo, Xiu Duan, Weizhen Fang, Yiyuan Sun, Baoyin Chen. Life is a lonely and painful journey, but because of you, I find it bearable. Thank you to my friends during my undergraduate's study, Yang Li, Xiaoliang Liu, Zhen Li, Guannan Dong, Yuxuan Xue, Bing Han. Thanks to my lifelong friends since high school: Hang Xun, Shang Liu. Thank you for being by my side through all the years of adulthood. Thanks to my hometown friends: Wenping Liu, Siqi Liu, Yong Guo, Jia Liu, Lanlan Wei. It is because of your presence that my hometown feels specific and warm.

I want to thank my family. My father, Jingquan Luo, you and mother raised me together. Mother gave me an optimistic outlook on life, and you taught me being a responsible person. If I am still considered an upright and kind person, it is thanks to both of you, and I can never repay this kindness. I would like to thank my grandfather, Shouzhen Luo, and my grandmother and my grandfather. My grandfather passed away during my second year of doctoral studies. I hope I can become the person you expect me to be, a grandson you can be proud of. Thank you to my sister, Xuanxuan Luo, and my brother-in-law, Zongchun Guo. To my lovely niece, Ziqing Guo, and nephew, Jinmu Guo, I am willing to devote the rest of my life to you. Thank you to my uncle and aunt, Jingen Luo and Xia Zhang. Special thanks to my best friend and cousin, Xiaotong Luo. You have always been by my side, giving me support. Thank you to my girlfriend, Jingwei Zhu, for accompanying me through countless wonderful moments, especially during this special time this year.

Finally, I want to thank every hardworking person in this world, every kind and upright person, every father and mother, every son and daughter. Your love is a rebellion against an absurd world. Life has no inherent meaning, but as I write this, realizing there are so many names I can thank, I understand the meaning of my remaining days.

This doctoral dissertation began in a cafe in Yanzhou, China, was defended at the Technical University of Munich in Germany, and was completed at the Hayden Library at MIT: this acknowledgment.

Peng Luo (罗鹏)

Hayden Library, MIT

Cambridge, USA

10.04.2024

Contents

Abstract	ii
Zusammenfassung.....	iii
List of Figures	ix
1. Introduction.....	1
1.1 Motivation.....	1
1.2 Objectives.....	5
1.3 Thesis structure.....	6
2. Theoretical background and related work	7
2.1 The nature of spatial data and approach to describe them.....	7
2.1.1 Spatial heterogeneity.....	8
2.1.2 Spatial dependence	8
2.1.3 Distance decay	9
2.2 Current progress in spatial analysis through the description of spatial data	9
2.2.1 Spatial patterns in spatial interaction data	9
2.2.2 Detection and explanation of spatial relationship between geographical variables.....	12
2.2.3 Spatial interpolation and spatial extrapolation.....	12
2.3 The limitations of current methods describing spatial data	13
2.3.1 Describing spatial interaction and spatial pattern in non-geospatial space.....	14
2.3.2 Explanation of spatial correlation in complex and non-linear interactions	15
2.3.3 Prediction of spatial distribution using sparse and biased samples.....	16
3 Summary of the work.....	18
3.1 Detection of overlapping spatial communities in spatial interaction data.....	18
3.2 Explanation of Non-linear interactions between variables	20
3.2.1 Theoretical foundation and assumption.....	20
3.2.2 Models for explaining spatial non-linear interactions	22
3.3 Spatial prediction with biased and sparse data	27
4. Conclusion and outlook.....	30
4.1 Conclusion.....	30
4.2 Limitations and outlook.....	30
Bibliography.....	32
Appendix: Publications	37

List of Figures

Figure 1-1. The framework of spatial analysis, spanning the collection of spatial data, its description, and the construction of models for three types of spatial tasks

Figure 1-2. Three cases where spatial data has not been effectively modeled in three spatial tasks: a) descriptive task; b) explanatory tasks; c) predictive task

Figure 2-1. Spatial data can beyond Euclidean space

Figure 2-2. Spatial networks and communities: (a) Spatial network of geographical units; (b) Disjoint communities; (c) Overlapping communities

Figure 2-3. The statistic assumptions of traditional regression models are violated in spatial data

Figure 2-4. The global nonlinear relationship between y and x ($y = ax^2$) are incorrectly interpreted as the process spatial nonstationarity ($\hat{\beta} = ax$) by GWR

Figure 2-5. The snapshot of the London government website, showing that London is establishing the world's largest air quality monitoring network.

Figure 3-1. The illustration of how the overlapping community structure can generate spatial interaction

Figure 3-2. The framework of exploring the overlapping community structure from spatial interaction data (i.e. human mobility)

Figure 3-3. The similar distribution of DEM and precipitation indicates their relationship

Figure 3-4. The progress of models proposed in this thesis for explaining non-linear interactions between geographical variables

Figure 3-5. The illustration of the GOZH model

Figure 3-6. The computational workflow of the LESH model. The steps (a-c) represent the process of calculating OPD, while steps (a, b, d, e) depict the process of calculating the SPD

Figure 3-7. The framework of GPI model

Figure 3-8. The framework of GHM

Figure 3-9. The framework of estimating crime risk based on the anomaly detection

1. Introduction

1.1 Motivation

The unique nature of space makes Geographic Information Science a valuable field of study (Longley et al. 2015; Liu et al. 2023), distinguishing geospatial data from other data types. Geospatial data usually contain observations of geographic variables at different locations and interaction intensity between locations. The uniqueness of geospatial data compared to other data types includes: 1) Communities distributed across space offer opportunities for complementarity, leading to *spatial interactions* between different locations (O’Kelly 2009); 2) Data collected from different locations, whether pertaining to a single or multiple geographical variables, typically exhibit *complex nonlinear interactions* and *are not independent* (Z. Li 2022); 3) Geospatial data are sparse samples from the *uneven* geographic distribution, which can hardly reveal universal patterns over space, and exhibit biases both in spatial and statistical distribution (OLIVER and WEBSTER 1990; Kwan 2012).

Overlooking the uniqueness of spatial data can lead to the incorporation of misleading information into analytical models, thus impairing their performance. It can result in deceptive conclusions and might even trigger ethical issues due to the inaccurate or biased findings (Chang 2021). Consequently, a thorough understanding of the characteristics of geospatial data is essential for precise spatial modeling and informed decision-making in spatial analysis and its associated disciplines.

The intricate nature of geospatial data presents challenges for modeling, making traditional statistical methods difficult to apply directly. On the positive side, these unique characteristics also bring valuable information and opportunities to spatial analysis. Through a deep understanding of geographic data, scientists have developed a range of methods for describing the uniqueness of spatial data, thereby developing appropriate models for various types of spatial analysis tasks. The spatial data can be described by the well-known spatial effects, such as spatial heterogeneity (Fotheringham, Brunson, and Charlton 2003), spatial dependence (Anselin 1995), and distance decay (Fotheringham 1981). Spatially explicit models are developed by incorporating these spatial effects to improve model performance or to minimize their side effect on the spatial task.

Understanding spatial data and providing a reasonable description are crucial for designing successful spatial models for various analysis tasks. The mission of spatial analysis is to describe geographical phenomena, explain their operational principles, and make reasonable predictions about these phenomena's characteristics in unknown areas or future times (Cressie 1990; Fotheringham and Rogerson 2008). The descriptive task focuses on the reflection of facts of spatial distribution or WHAT spatial pattern exists. The

explanatory task goes deeper into the relationships between values from different locations or between variables, or WHY they are correlated. Predictive tasks involve constructing appropriate models based on identified spatial patterns and interpreted relationships. These models are then used to predict values for unobserved areas or future time periods based on collected geographical data. After the spatial data has been collected as the representation for the geographical world, the method of describing spatial data fundamentally influences various spatial analysis tasks, including descriptive tasks, explanatory tasks, and predictive tasks (Figure 1-1).

First, descriptive tasks involve identifying spatial patterns and structures, including quantifying the degree of spatial clustering of variables, discovering potential topological and semantic structures in space. Spatial data can be modeled in different forms include field models, object models, and network models, depending on the purpose of data description and the method of data acquisition (Michael F. Goodchild 1992). For describing the degree of clustering and dispersion of spatial distribution, field models and object models are often chosen. The spatial pattern is characterized by how individual entities are positioned in space and the geographic connections between them (Chou 1995). Methods such as the Getis-Ord G_i^* for identifying hotspots and cold spots (Getis and Ord 1992), and Moran's I index for describing spatial autocorrelation (Moran 1950), are pivotal in recognizing spatial patterns. To uncover the structural information of geospatial areas, we often use geographic data with interaction or connection information, such as static road networks and dynamic human mobility data, where geographic data can be expressed as networks or graphs connecting locations. A key aspect of describing geographic interaction data is estimating the strength of spatial interactions between different locations. For example, a widely accepted concept is distance decay (Fotheringham 1981), indicating that the interaction between two locations decreases as the distance between them increases.

Second, explaining the relationships between geographical variables is a core mission of geography (Fotheringham, Brunson, and Charlton 2003). Geographers aim to understand the correlations between geographic variables, especially since such correlations often vary across space (Fotheringham and Li 2023). This widespread variation in the distribution and relationships of geographic variables in space is described as spatial heterogeneity (Anselin 2010). To explain the associations between geographic variables under the premise of spatial heterogeneity, common methods include adding parameters for spatial variation to the model, such as Geographically Weighted Regression (GWR) and Bayesian spatially varying coefficient models (Fotheringham, Brunson, and Charlton 2003; Gelfand et al. 2003).

Third, for spatial prediction tasks, the understanding of spatial data features and the accuracy of their description significantly affect the performance of spatial predictions. The values of geographic variables

at different locations often have potential dependencies, described by spatial autocorrelation (Chou 1995). The presence of spatial autocorrelation makes the effective sample size insufficient in spatial prediction. Spatial sample points are often sparse and traditional prediction models may face underfitting problem. However, by utilizing spatial autocorrelation, we can model relationships between sample points collected from different locations, thus making predictions in unknown areas. For example, in geostatistics, unbiased estimation is achieved through constructing semivariograms based on known sample points (Luo and Song 2021; Luo et al. 2023).

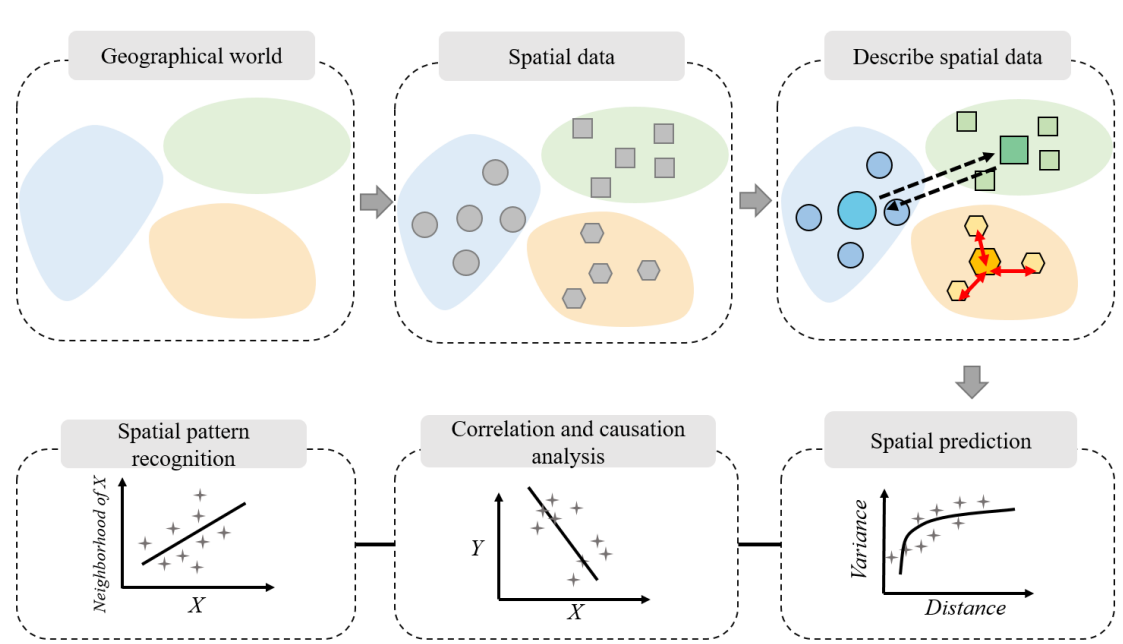


Figure 1-1. The framework of spatial analysis, spanning the collection of spatial data, its description, and the construction of models for three types of spatial tasks.

Despite the development of numerous spatial analysis methods through reasonable description of spatial data, many characteristics of spatial data remain inadequately described and understood, thus limiting the performance and applicability of current spatial analysis methods. This thesis focus on three aspects of spatial data that have not been addressed:

- 1) Spatial interactions may arise from overlapping community structures (Luo and Zhu 2022). There is the challenge of mining overlapping community structures from spatial interaction data (Figure 1-2a).
- 2) The values of geographic variables often have weak statistical assumptions, with nonlinear interactions between different geographic variables (Luo et al. 2022; Z. Li 2022; Fotheringham and Li 2023). As is

shown in Figure 1-2b, the confounding effect may exist, that the relationship between X_1 and Y varies depending on whether the value of another variable, X_2 , is high or low.

3) The issue of spatial data being sparse in space can lead to biased information provided (Figure 1-2c), both in terms of statistical distribution and spatial distribution (Michael F. Goodchild 1989). In spatial interpolation, building a prediction model, such as Kriging, over the global space often fails to accurately capture data attributes on non-stationary surfaces. The stratified modeling strategy, for example, training models separately in sub-regions, is widely used. However, this approach can lead to underfitting because the number of samples in each sub-region may be limited due to sparse sampling issues. In addition, In areas of high heterogeneity, sparse spatial sampling can lead to inaccuracies in capturing the correct statistical distribution, potentially skewing the data towards characteristics of sub-regions with more sampling points.

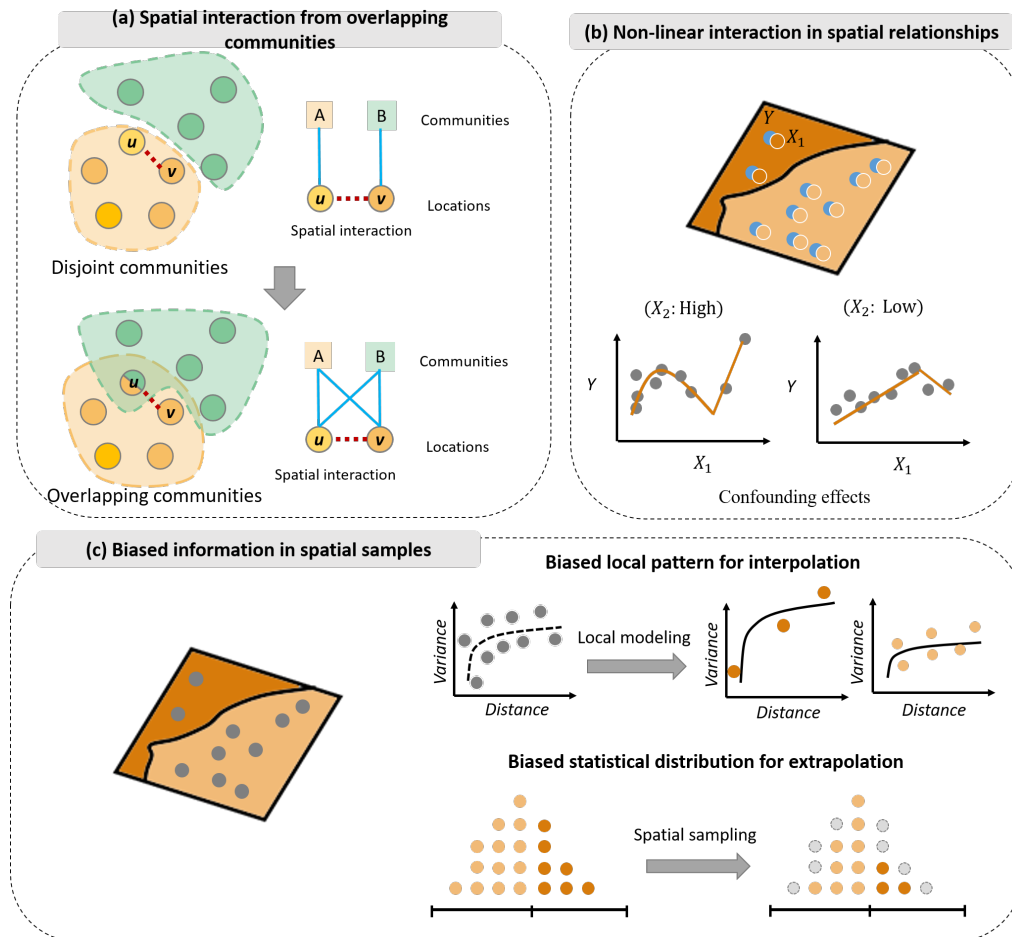


Figure 1-2. Three cases where spatial data has not been effectively modeled in three spatial tasks: a) descriptive task; b) explanatory tasks; c) predictive task

In fact, the methods of describing spatial data, specifically spatial effects and associations, extend far beyond what is currently proposed, such as dependence and heterogeneity. Numerous spatial associations can be characterized. For instance, the generative mechanisms of spatial interactions and communities, the similarity in spatial patterns of two geographic variables can indicate their correlation to a certain extent, and the spatial relationships across different regions and scales can be collaboratively modeled and predicted. To better describe spatial data, a more generalized modeling of spatial associations is required.

In this thesis, we aim to gain a thorough and integrated understanding of spatial data's uniqueness. This will enable more generalized modeling of spatial associations and lead to more accurate spatial analysis.

1.1 Objectives

The primary aim of this research is to model spatial relationships, aiming to provide a more accurate description of spatial data and thereby enhance spatial analysis tasks. In this thesis, we define this approach as *generalized spatial association modeling*. Our goal is to detect the comprehensive nature of spatial data. The main characteristic of spatial data that introduces challenges and potential benefits to spatial models is that data from different locations are potentially associated in some way. We believe that current methods used to describe spatial data, such as spatial heterogeneity and dependence, also aim to detect spatial associations, though from different perspectives.

In this thesis, we aim to identify generalized spatial associations among spatial data and focus on addressing the following research questions:

RQ-1: Can potential overlapping community structures in space be identified using spatial interaction data?

RQ-2: How can nonlinear interactions between geographical variables be identified under the premise of weak assumptions?

RQ-3: How can accurate spatial predictions be made in cases of sparse and biased samples?

By modeling spatial associations in a more generalized manner, this work aims to achieve three objectives:

O-1: To uncover the overlapping community structures that drive spatial interactions;

O-2: To investigate nonlinear interactions between geographical variables under weak statistical assumptions, while also elucidating the contribution of single variables in the context of multivariate interactions;

O-3: To explore suitable methods for spatial prediction, including spatial interpolation and extrapolation, in the presence of sparse and biased samples.

1.2 Thesis structure

This is a cumulative thesis, comprising six peer-reviewed journal papers and one peer-reviewed conference paper. Following this chapter, Chapter 2 is dedicated to the current research on spatial data and spatial analysis methods. It begins with the nature of spatial data and spatial effects to describe spatial data. It then summarizes the research progress in three spatial analysis tasks: description, explanation, and prediction. Chapter 3 provides an summary of the conducted research. Given the nature of spatial data, the publications included in this thesis aim to improve the performance of spatial analysis through more comprehensive modeling of spatial association. They cover the three aforementioned spatial analysis tasks. Chapter 4 concludes the thesis with the main findings of the research, its shortcomings and limitations, and potential future directions for expansion.

Part I: Detection of overlapping spatial communities in spatial interaction data

1. **(A1) Luo, P.** and Zhu, D., 2022, November. Sensing overlapping geospatial communities from human movements using graph affiliation generation models. *In Proceedings of the 5th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery* (pp. 1-9).

Part II: Explanation of Non-linear interactions between variables

2. **(A2) Luo, P.**, Song, Y. and Wu, P., 2021. Spatial disparities in trade-offs: economic and environmental impacts of road infrastructure on continental level. *GIScience and Remote Sensing*, 58(5), pp.756-775.
3. **(A3) Luo, P.**, Song, Y., Huang, X., Ma, H., Liu, J., Yao, Y. and Meng, L., 2022. Identifying determinants of spatio-temporal disparities in soil moisture of the Northern Hemisphere using a geographically optimal zones-based heterogeneity model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 185, pp.111-128
4. **(A4) Li, Y.[#], Luo, P.[#]**(co-first), Song, Y., Zhang, L., Qu, Y and Hou, Z., 2023. A locally explained heterogeneity model for examining wetland disparity. *International Journal of Digital Earth*, 13(2), p.4533-4552.
5. **(A5) Luo, P.**, et al. (2023). Measuring univariate effects in the interaction of geographical patterns. *International Journal of Geographical Information Science*. (Under Review)

Part III: Prediction of spatial distribution using sparse and biased samples

6. (A6) **Luo, P.**, Song, Y., Zhu, D., Cheng J. and Meng L. A Generalized Spatial Heterogeneity Model for Interpolation.2022 *International Journal of Geographical Information Science*, 37(3), 634-659
7. (A7) Yao, Y., Dong, A., Liu, Z., Jiang, Y., Guo, Z., Cheng, J., Guan, Q. and **Luo, P***. (correspondence, project lead), 2023. Extracting the pickpocketing information implied in the built environment by treating it as the anomalies. *Cities*, 143, p.104575.

1. Theoretical background and related work

In this section, we first discuss the characteristics of spatial data and the methods used to describe spatial data, focusing on several spatial effects. Second, we summarize the research progress into three tasks of spatial analysis. Finally, we elucidate the limitations corresponding to each of the three spatial tasks.

2.1 The nature of spatial data and approach to describe them

Geographic data are from the uneven sampling of geographic variables in space. The spatial distribution characteristics of geographic variables are directly or indirectly reflected in the spatial data:

1. Spatially uneven numerical distribution. This results from either the heterogeneous distribution of numerical values of geographical variables across space or from inappropriate spatial sampling (J.-F. Wang et al. 2012).
2. Spatially uneven data-generating process, leading to geographical variable relationships that differ across different areas (Anselin 1988).
3. Data from different locations have relationships. This may from the generation process of geographic variables, including geographic diffusion processes (Chin et al. 2017) and spatial spillovers (Capello 2009). Depending on the type of spatial data, the relationship between different areas can be numerical proximity or some form of spatial interaction with varying intensities (O’Kelly 2009).

Based on the characteristics of spatial data, several concepts have been developed to describe spatial data, which can also be referred to as spatial effects. These are: spatial heterogeneity, spatial dependence, and distance decay.

2.1.1 Spatial heterogeneity

Spatial heterogeneity suggests that the value/form of a geographical variable or the underlying process differs over space (De Marsily et al. 2005; Fotheringham, Brunson, and Charlton 2003; Getis and Ord 1992). Spatial heterogeneity in value/ form is a common occurrence, which can be divided into local heterogeneity and stratified heterogeneity (J. Guo et al. 2022). The heterogeneity of spatial processes refers to the differences in the generating processing of data at different locations. At the level of algorithms and models, this heterogeneity is manifested in the form and parameters of models exhibiting spatial variation, also known as spatial non-stationarity (Liu et al. 2023). This variation often leads to geographical phenomena displaying unique characteristics or patterns in distinct areas (Luo and Song 2021).

To capture spatial heterogeneity, models can be structured either discretely or continuously. When considering discrete heterogeneity, multi-level models and spatial regimes are frequently applied (Fotheringham and Li 2023; Anselin and Amaral 2023). On the other hand, models like GWR and Spatial Eigenvector Filtering are employed for continuous heterogeneity, allowing for the analysis of spatially varying coefficients (Fotheringham, Brunson, and Charlton 2003; Griffith 2003).

2.1.2 Spatial dependence

Spatial dependence refers to the correlation that exists between neighboring locations in geospatial space, which reflects the tendency of geographical phenomena to cluster or disperse in space (Anselin 1995). The existence of spatial dependence forms the foundation of geographical analysis. It indicates that the assumption of variable independence, typically upheld in classical statistics, is often violated in geospatial data, wherein the count of independent sample points is fewer than the total number of samples (Griffith 2005).

A common method to describe spatial dependence is spatial autocorrelation. Methods for assessing spatial autocorrelation can be divided into global autocorrelation indices and local autocorrelation indices. Global autocorrelation indices include Moran's I index and Geary's C coefficient, among others. Local autocorrelation indices are used to quantify spatial dependence in local spaces. For example, Anselin introduced the local indicator for spatial association (Anselin 1995), and Getis proposed the generalized G statistic.

Additionally, the semivariogram from geostatistics can also be used to assess spatial dependence (Cressie 1990; Luo and Song 2021; Luo et al. 2023). When the value of the semivariogram increases with distance until reaching a stable value (known as the sill), it indicates spatial dependence within that range of distance;

the sample points are spatially related. If the semivariogram value rapidly increases to the sill, this suggests that spatial dependence significantly decreases within a short distance, indicating greater spatial variability. Conversely, if the increase in the semivariogram is gradual, it suggests that even at greater distances, the sample points maintain a certain level of similarity, indicating stronger spatial dependence and less spatial heterogeneity.

It should be noted that spatial dependence and spatial heterogeneity are often represent the same spatial distribution, but from various aspects, so it can be difficult to distinguish from one another. This inverse problem stems from the challenge of identifying the processes that generate such spatial patterns based on available spatial data (Anselin 2010).

2.1.3 Distance decay

In geographic spaces, there exist varying intensities of connections between features at different locations, which move and exchange in diverse ways (Liu et al. 2023). This process is known as spatial interaction (O’Kelly 2009). As defined for spatial data, an essential component is the information that records the spatial interactions between different locations. A widely accepted method to describe spatial interaction is the principle of distance decay, suggesting that the closer two locations are in spatial distance, the higher the potential intensity of their interactions (S. Gao et al. 2013). The distance decay can be modeled from various function, such as gravity model (Chen 2015).

However, with technological advancements, the cost of interactions in both physical and virtual spaces has significantly decreased (Luo and Zhu 2022). Different locations are closely connected by dense population flows, or individuals can interact through social media despite being far apart. Therefore, the importance of geographic distance needs to be reassessed (Liu et al. 2023). Despite this, extensive research indicates that geographic constraints still exist within highly mobile spatial networks, making distance remain an important factor in modeling spatial interactions.

2.2 Current progress in spatial analysis through the description of spatial data

2.2.1 Spatial patterns in spatial interaction data

Describing the spatial distribution patterns of geographic data with spatial interactions is one of the research focuses in geography (Liu et al. 2023). With the advent of the big data era, an increasing amount of data involving spatial interactions is being captured, such as population movement trajectories, taxi origin-destination (OD) data, and trade networks (Figure 2-1). This type of data is characterized by its

representation as a spatial network composed of nodes and edges (Luo and Zhu 2022). Here, nodes represent different locations, and the edges connecting these nodes are weighted to record the strength of interactions between two locations.

Spatial interactions arise from differences between locations, including complementarity, intervening opportunities, and transferability (Liu et al. 2020). Therefore, geographic patterns manifest as spatial differentiation with distinct attributes, suggesting that spatial distributions can be characterized by community structures with unique features, which are fundamental to the formation of spatial interactions. Locations within the same community exhibit stronger spatial interaction strengths, whereas interactions between locations in different communities are weaker. Different communities offer services and opportunities of varying functional types, often presenting degrees of irreplaceability in space. Hence, complementarity exists between different communities, leading to the movement and interaction of people in space.

It is possible to infer geographic community structures from spatial interaction data, thanks to the wide availability of spatial interaction data and the close link between spatial interactions and spatial community structures (Liu et al. 2020; Jia et al. 2022). The task of dividing spatial units into non-overlapping community areas based on spatial interaction data is defined as community detection (Hong and Yao 2019). Geographic community detection refers to the process of identifying clusters of nodes with tightly connected characteristics in networks that include geographic location information. Geographic community detection aims to describe the spatial patterns of a spatial network based on spatial interaction data, which involves identifying and understanding the network's structural and functional patterns. For geographic spatial networks, uncovering spatial patterns can reveal regular information about geographic variables in space: 1) understanding the function of locations within the entire network and identifying key locations; 2) understanding the propagation characteristics of geographic variables in space; 3) facilitating personalized urban management and services.

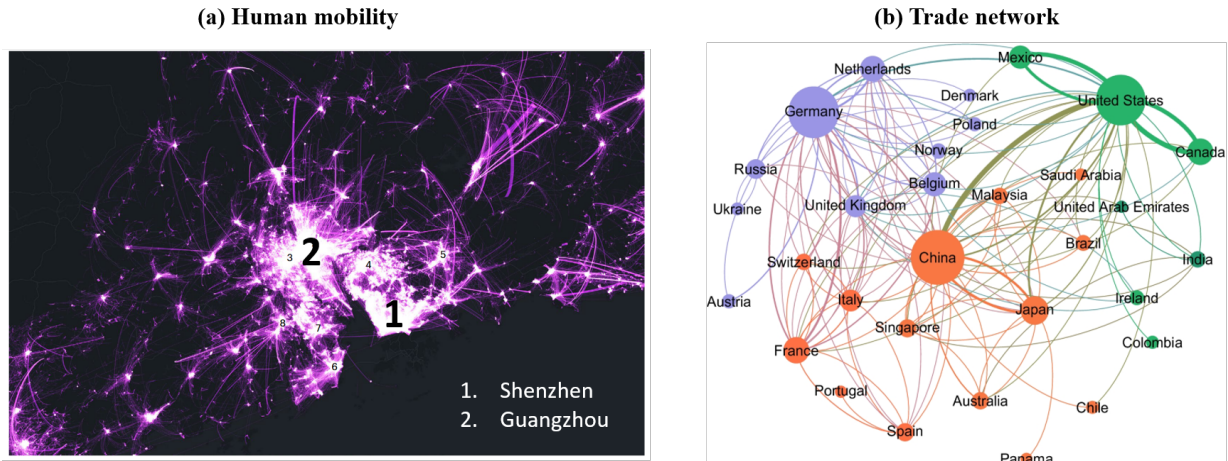


Figure 2-1. Spatial data can beyond Euclidean space. (adopted from (Dong et al. 2024) and (Setayesh, Sourati Hassan Zadeh, and Bahrak 2022))

For geographic community detection, a key element is accurately quantifying the spatial interaction strength between two nodes in space (Y. Wang et al. 2020; M. Li et al. 2021). There are several approaches to this:

1. Geographical distance: The principle of setting spatial interaction strength on this basis is derived from Tobler's first law of geography, which posits that distance determines the similarity between two places. This distance could be Euclidean distance or road distance. For example, (Hong and Yao 2019) based their weighting of the spatial network on the distance between two road nodes to perform hierarchical community detection in urban road networks.

2. Mobility interaction of geographic entities: This method involves using the intensity of interactions between places, such as human movement (D. Guo et al. 2018), e-scooter flows, and call flow between two cells (S. Gao et al. 2013). Since mobility data can capture time-dynamic information, this approach can also facilitate dynamic community detection (Jia et al. 2022).

3. Social intensity: This method considers the social intensity between two geographic entities as the spatial interaction intensity. For instance, (Yao et al. 2021) used the perception assessment from street view images between two road segments (e.g., wealth, safety) as the measure of spatial interaction intensity.

Each of these methods offers a unique perspective on understanding and measuring spatial interactions, contributing to the detection of geographic communities by considering both physical proximity and the intensity of social and physical interactions.

Conventional methods for detecting communities primarily investigate communities by examining the structure of networks, identifying disjoint communities, such as through methods of partitioning graphs (Newman 2013), statistical inference (Zhao et al. 2017), and spectral clustering (Zhou and Amini 2018). For instance, aiming to enhance modularity, the Louvain algorithm was developed for rapid identification of communities and has seen extensive application (Blondel et al. 2008).

2.2.2 Detection and explanation of spatial relationship between geographical variables

In geographical analysis, a primary objective is to understand the processes behind geographical data and explain how geographical variables are interrelated (Fotheringham, Brunson, and Charlton 2003). Grasping these connections is crucial for mastering the mechanisms generating the data, predicting future trends, and guiding policy and decision-making. Through statistical methods designed specifically for geographical data, one can explore the quantitative relationships between geographical variables. Traditional statistical models often focus on non-spatial correlations, assuming a uniform relationship across space, with model parameters being constant throughout the area (Z. Li 2022). Initially, global linear regression models produce a single regression equation from data generated at different locations, assuming that the data generation process within the study area is the same. Second, global spatial regression methods have been developed to consider spatial effects by explicitly incorporating spatial autocorrelation into the regression models, such as spatial lag models and spatial error models (Anselin 1988; 1992). However, their model parameters remain constant in space (Haining 1990). These assumptions of parameter spatial stability in global models can lead to biased results and incorrect assumptions.

The relationships between different variables generally vary across space, necessitating local parameter estimation. Local models have been developed to address this issue, allowing model parameters to vary with location. Examples of local models include GWR (Fotheringham, Brunson, and Charlton 2003), Bayesian Spatially Varying Coefficient Models (Gelfand et al. 2003) and Eigenvector Spatial Filtering Models (Griffith 2003; Tiefelsdorf and Griffith 2007).

2.2.3 Spatial interpolation and spatial extrapolation

The essence of spatial prediction lies in utilizing observational data from known locations to estimate geographic variable values in uncharted territories, extending possibly to a global distribution (Mitas and Mitasova, n.d.; Zhu and Cao 2023). The classification of spatial prediction methodologies bifurcates into interpolation and extrapolation. Spatial interpolation is a method for predicting the values of unknown data points situated between known data points (Cressie 1990). In geospatial analysis, this method is predominantly employed to estimate values at locations within a region based on a set of known values

from that same area. Spatial extrapolation refers to predicting the values of data points beyond the areas where known data are collected. This range may encompass unsampled areas within a study region or entirely new study areas (Zhu et al. 2022).

Spatial interpolation includes deterministic methods and statistical methods. Deterministic methods such as Voronoi natural neighbor interpolation, Inverse Distance Weighting (IDW), and Triangular Irregular Networks (TIN), rely on predefined spatial relationships to structure geospatial phenomenon predictions (Lam 1983). Statistical methods, mainly geostatistics, undertake an extensive learning process to model geographic data as realizations of spatial processes. Geostatistics focuses on depicting spatial variability using semi-variograms and predicts unobserved values with the kriging method series (OLIVER and WEBSTER 1990).

For spatial extrapolation, the general approach involves constructing relationships between explanatory variables and variables to be predicted based on existing data from a study area, then generalizing these relationships to areas without samples for prediction. Regression models are the most commonly utilized methods for extrapolation. Non-spatial regression models include univariate linear regression, multivariate linear regression, nonlinear models, etc. Moreover, machine learning regression, such as random forest, has been increasingly applied to spatial prediction in recent years (Yao et al. 2023). Due to their robust data-fitting capabilities and ability to capture non-linearities, they often outperform traditional regression models (Z. Li 2022).

Spatial regression models primarily estimate model parameters based on existing data, serving as explanatory models rather than predictive ones. However, in some instances, they can also be used for predictive extrapolation. Spatial regression formalizes spatial relationships as correlation structures within a linear regression model framework (Anselin 1988). This involves specifying, estimating, and diagnosing regression models that incorporate spatial effects, thus optimizing the model for spatial prediction. Spatial regression models are designed to manage the complexities of spatial data, including spatial autocorrelation and heterogeneity, by directly embedding spatial dependence into the regression analysis. This not only aids in understanding spatial dynamics but also improves prediction accuracy by considering the spatial relationships among variables (Anselin 1992).

2.3 The limitations of current methods describing spatial data

2.3.1 Describing spatial interaction and spatial pattern in non-geospatial space

Traditional descriptions of spatial patterns based on spatial interaction strength come with significant constraints, assuming that a location belongs to only one community with homogeneous characteristics (Figure 2-2b). However, it's possible for a location to overlap multiple communities with distinct attributes (Figure 2-2c). For example, social services such as police deployment and medical services are often optimized as community-based resource allocations. Residents living between two clinics might have access to both, positioning them in an overlapping service area of the two clinics. Assigning residents exclusively to one clinic under a non-overlapping community division could lead to wastage of public resources.

Moreover, according to Granovetter's theory (S. Granovetter 1973), connections within a network can be complex, classified into strong and weak ties. Structurally embedded (tightly connected) edges are often strong socially, while remote edges spanning different network sections tend to be weak socially. Most existing community detection algorithms presuppose that a set of nodes constitutes a community only if they are connected more strongly than expected. This fundamental assumption overlooks the fact that nodes with weak ties can form communities that offer greater informational benefits. Often, it is acquaintances, not friends, who provide us with information beyond our immediate sphere. Long-distance and weak connections form critical community structures that are indispensable for societal benefits.

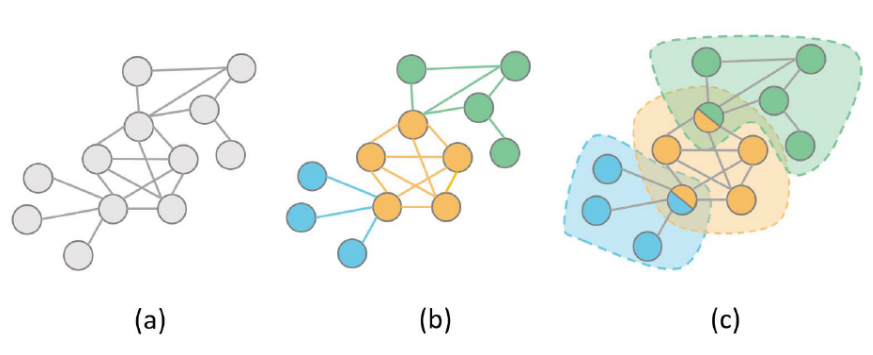


Figure 2-2. Spatial networks and communities: (a) Spatial network of geographical units; (b) Disjoint communities; (c) Overlapping communities

Therefore, considering that communities may overlap spatially and that the strength of connections between community members varies, inferring such complex overlapping community structures from spatial interaction networks is particularly important. Achieving this goal will help optimize resource allocation, enhance the efficiency of public services, and strengthen community cohesion and social capital.

2.3.2 Explanation of spatial correlation in complex and non-linear interactions

Most prevalent spatial correlation analysis methods are grounded in the linear regression paradigm, integrating spatial parameters. This adherence results in the statistical assumptions conforming to the linear paradigm, characterized by:

- Normal distribution
- Independent and identically distributed (i.i.d.) data
- Linear relationships among variables

In contrast, geographical data often present:

- Unbounded distributions, challenging the traditional statistical assumptions due to the absence of i.i.d. and normal distribution.
- Pronounced non-linear interrelations among variables, complicating the modeling of spatial data using traditional linear paradigm-based methods.

As is shown in Figure 2-3, the average house price data in the UK has been found to exhibit strong spatial dependence, particularly evident in regions such as London. Additionally, there is a pronounced skewness in the overall distribution of house prices, with significant disparities in statistical distributions across different regions.

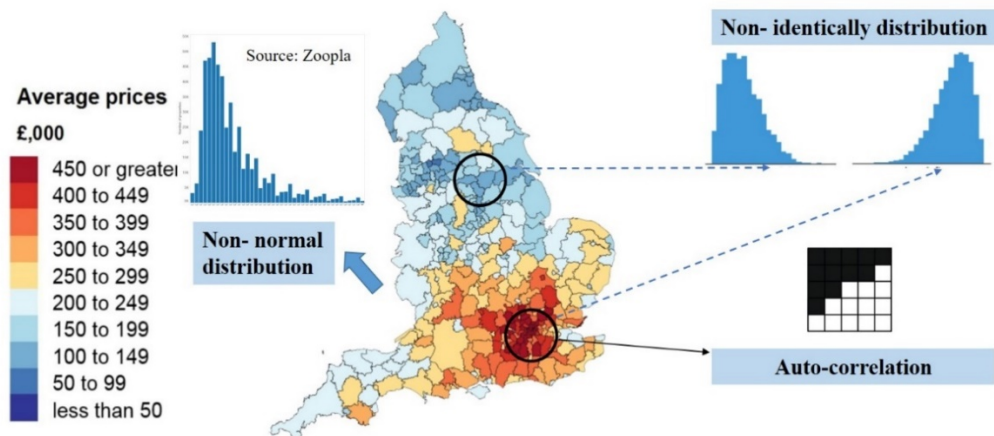


Figure 2-3. The statistic assumptions of traditional regression models are violated in spatial data

The statistical characteristics of spatial data can lead to incorrect inferences with traditional methods, especially those that utilize the linear regression paradigm. Figure 2-4 demonstrates the problem of capturing nonlinear relationships using GWR (Sachdeva et al. 2022). When explanatory variables and response variables exhibit nonlinear relationships, even if these relationships are stable across space, GWR may inaccurately interpret this as spatial variation. In summary, it is crucial to address the challenges posed by the unconstrained distribution of spatial data, as well as the complex nonlinear interactions between different variables for identifying spatial correlations.

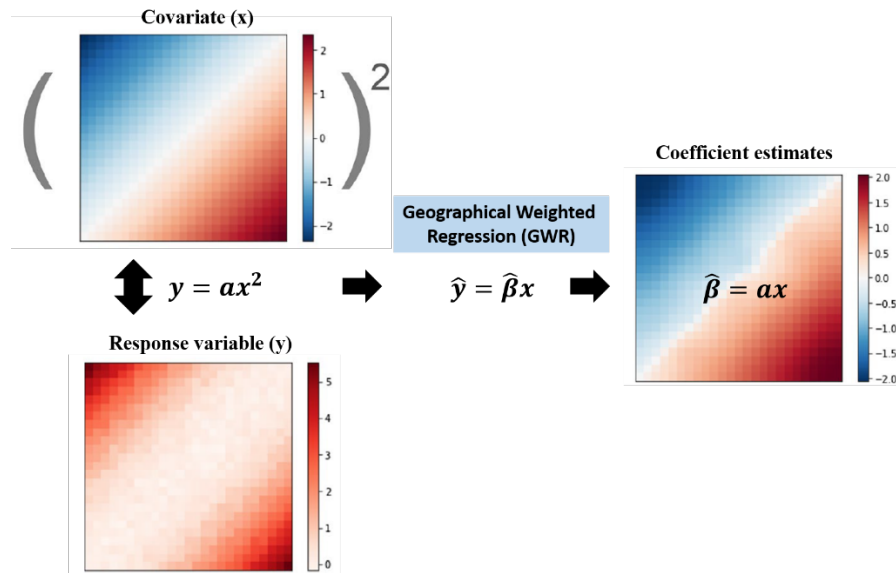


Figure 2-4. The global nonlinear relationship between y and x ($y = ax^2$) are incorrectly interpreted as the process spatial nonstationarity ($\hat{\beta} = ax$) by GWR (adopted from Sachdeva et al. 2022).

2.3.3 Prediction of spatial distribution using sparse and biased samples

One of the main challenge for the current spatial prediction models is addressing sparse and biased samples. Spatial data are samples from the real geographic world (M F Goodchild, Anselin, and Deichmann 1993). Limited by time and economic costs, the distribution of spatial data often exhibits sparsity, and the actual obtained geographic spatial distribution data is usually incomplete (A.S. and B.K. 1991). Moreover, due to the heterogeneity of spatial data itself, no sampling method can perfectly represent the distribution of geographical variables over space, hence spatial data inherently possess uncertainty. A common case of spatial prediction is predicting the distribution of air quality based on in-situ sensors. Due to sampling cost constraints, the sensors we can deploy are often very sparse. In the case illustrated in Figure 2-5, we rely on a spatial quality monitoring network of around 100 sites to obtain an assessment of air quality throughout the city of London. Given the limited samples and size of the spatial prediction area, this poses a major

challenge for spatial prediction models, whether spatial interpolation or extrapolation. Are 100 points sufficient to capture the entire spatial distribution of air quality in London? If not, predictions of air quality are likely to contain significant errors and uncertainties. If predictions for certain areas of air quality are overestimated or underestimated, it could lead to misleading policies. Who then bears responsibility for the local residents? Therefore, we must have a clear understanding of the predictive uncertainty of spatial forecasting in cases of sparse samples and endeavor to address it.

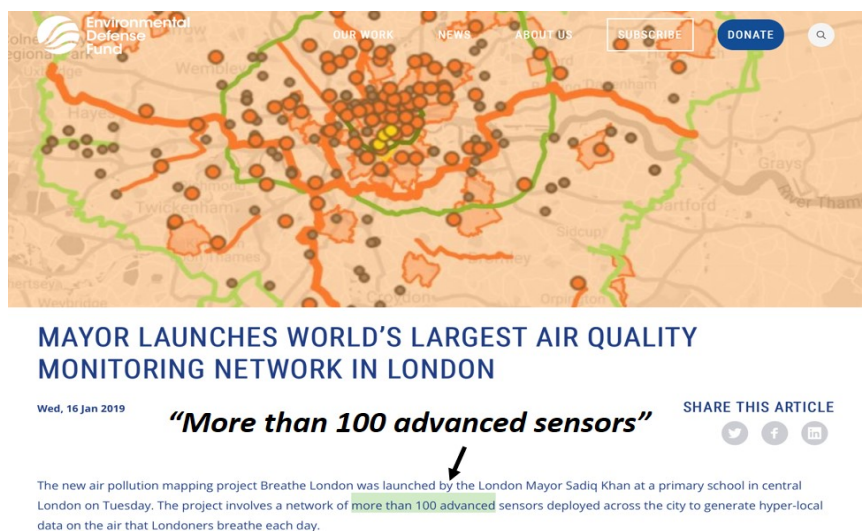


Figure 2-5. The snapshot of the London government website², showing that London is establishing the world's largest air quality monitoring network.

In the context of spatial interpolation, the incompleteness of sampling data hinders the ability to construct accurate spatial relationships, leading to imprecise estimates of spatial structures. Current models, when predicting spatial data, often rely on a strong assumption of spatial stationarity, they assume that the process generating spatial data is uniform across the study area (Luo et al. 2023). Traditional spatial interpolation methods, such as geostatistics, typically assume that variables exhibit second-order stationarity. Similarly, when utilizing deep learning models—transforming spatial prediction tasks into computer vision tasks using convolutional neural networks—the inherent mechanism of convolutional kernels implies an assumption of distributional stationarity. However, for the large area, capturing spatial non-stationarity over the entire space becomes challenging. This indicates that within a study area, more than one data generation process may exist (B. Gao et al. 2020). A common solution to this issue is spatial zoning for modeling purposes. Nevertheless, in cases of sparse sampling, such zoning inevitably results in a very limited number

² <https://www.london.gov.uk/press-releases/mayoral/to-identify-londons-toxic-air-hotspots>

of observable points available for modeling within each zone, leading to the construction of inaccurate semivariograms, and thus, diminished predictive performance (Luo et al. 2023).

For spatial extrapolation prediction, the relationship between explanatory variables and variables to be predicted is constructed based on data collected from known regions. Due to the existence of spatial heterogeneity, sparse samples often contain incomplete and biased information, leading to unreliable conclusions in the models constructed (Figure 1-2).

3 Summary of the work

3.1 Detection of overlapping spatial communities in spatial interaction data

Related publication: Luo, P. and Zhu, D., 2022, November. Sensing overlapping geospatial communities from human movements using graph affiliation generation models. In Proceedings of the 5th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery (pp. 1-9).

To mine overlapping spatial community structures from spatial interaction data, we introduce a hypothesis: the more communities two locations simultaneously belong to, the stronger their spatial interaction (Luo and Zhu 2022; 2024). Thus, the intensity of spatial interactions can be estimated by the strength of the locations' affiliations to communities. As illustrated in Figure 3-1, assume there are two communities, Com_1 (colored red) and Com_2 (colored green), and three locations A , B , and C in a city. Each location has a certain degree of affiliation with both communities. For instance, A belongs exclusively to Com_1 , C to Com_2 , while B is at the overlapping of Com_1 and Com_2 . Our hypothesis posits that spatial interactions (e.g., human mobility) between two locations occur because they share at least one common community. Based on this hypothesis, there would be population movement between A and B , and between B and C , while direct movement between A and C might be less feasible and B would act as a intersections. In this way, the overlapping community structures can be used to reconstruct spatial interaction networks, offering an opportunity to infer community structures from the collected spatial interaction networks.

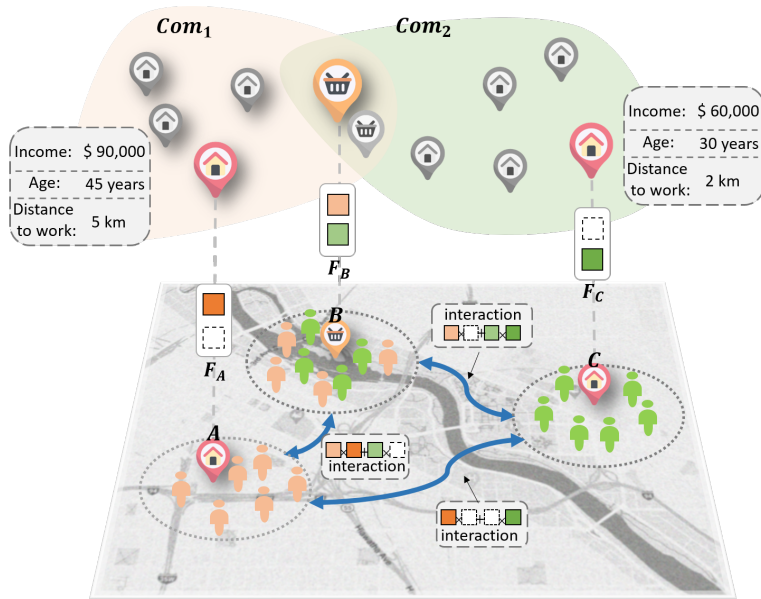


Figure 3-1. The illustration of how the overlapping community structure can generate spatial interaction (Luo and Zhu 2024)

Our task is to infer the spatial distribution of communities from available spatial interaction data (such as human mobility). We have adopted a graph generation approach to address this issue. Knowing that the strength of spatial interactions can be estimated through the affiliation strength of two nodes to different communities, we can initialize the affiliation strength of each node to the communities. Subsequently, we estimate the spatial interaction network and compare it with the actual spatial interaction network. The differences between these networks are used to optimize the affiliation strength. This optimization process can be facilitated using Graph Convolutional Networks (GCN). In this graph, the affiliation strength of each location to different communities is represented as node attributes, while the intensity of the spatial interaction network is represented as the edge weight between nodes.

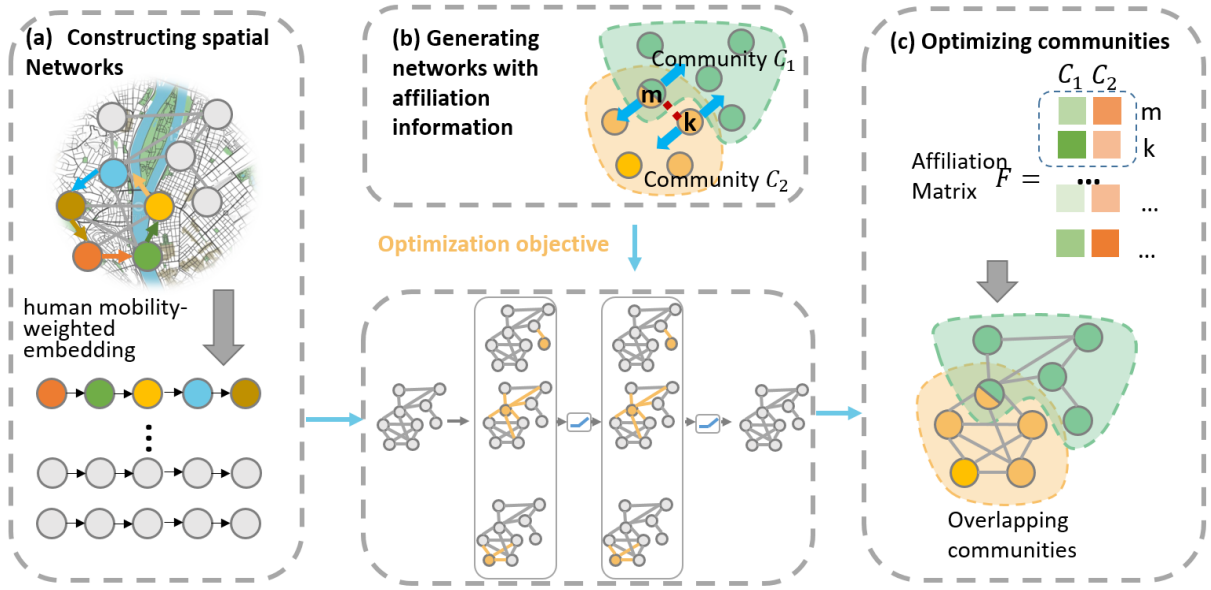


Figure 3-2. The framework of exploring the overlapping community structure from spatial interaction data (i.e. human mobility).

As shown in Figure 3-2, a) to better extract the features of each location, we first embed the local spatial information of each location using the node2vec model (Grover and Leskovec 2016). This embedded information serves as the initial attributes for each node; b) subsequently, we utilize a GCN model to estimate the edge weights based on these attributes (Kipf and Welling 2017). c) After calculating the weights between each pair of edges, we can compare these with the actual data to compute the loss, which then feeds back into the GCN for optimization. We conducted a case study using mobile positioning data from the Twin Cities Metropolitan Area in Minnesota to validate our model's effectiveness in real-world human mobility networks. Our empirical results revealed the overlapping spatial structure of communities, the overlapping intensity for each location, and the spatially heterogeneous structure of community affiliations in the Twin Cities.

3.2 Explanation of Non-linear interactions between variables

3.2.1 Theoretical foundation and assumption

Spatial relationships often exhibit nonlinearity and the variables involved are not independent but interact in complex ways. We introduce a novel approach to identify spatial relationships: the similarity of spatial patterns. As demonstrated in Figure 3-3, based on the spatial distribution of the Digital Elevation Model

(DEM) and Precipitation in the United States, which exhibit remarkably similar features, we intuitively infer a spatial relationship between them. Indeed, similar models, such as the spatial stratified heterogeneity (SSH) model, have been used to analyze such relationships.

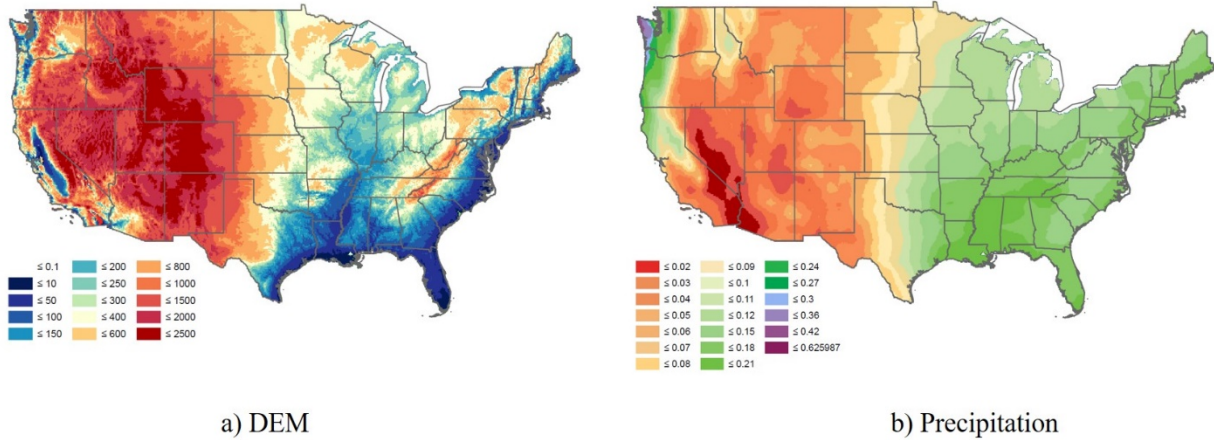


Figure 3-3. The similar distribution of DEM and precipitation indicates their relationship

Building on the concept of SSH, we have developed and expanded a theoretical and methodological framework for spatial correlation analysis based on the similarity of spatial distributions. The core concept of our proposed models is that the interactions between geographical patterns can reveal spatial associations. This model posits that the behavior of a response variable is shaped by the interactions among multiple explanatory variables. Specifically, the influence of an explanatory variable X on the spatial pattern of a response variable Y signifies the spatial association between X and Y . Spatial patterns can be defined in various ways, often demonstrating that a geographical variable's distribution forms distinct, relatively homogeneous subregions. Within these subregions, values are similar, while across different subregions, they are dissimilar.

In this thesis, we delved into the theory of spatially stratified heterogeneity, experimented with its application (A2), accuracy (A3), and interpretability (A4), and ultimately developed a model capable of estimating spatial variability relationships (A5).

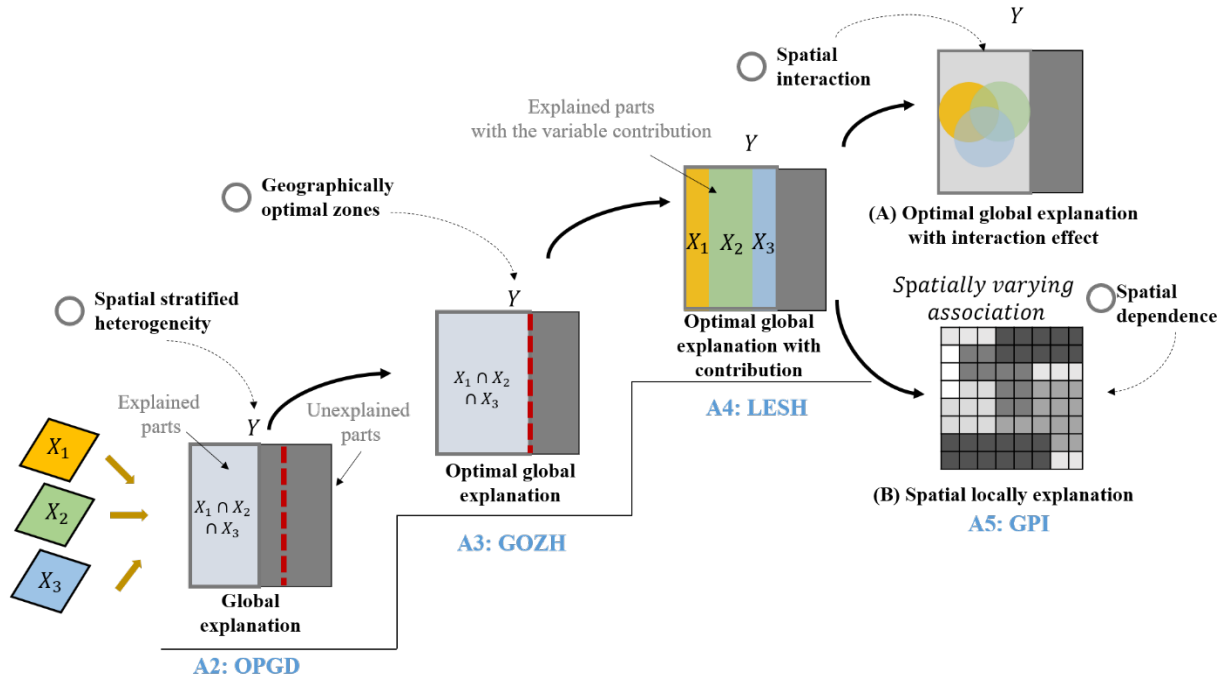


Figure 3-4. The progress of models proposed in this thesis for explaining non-linear interactions between geographical variables.

3.2.2 Models for explaining spatial non-linear interactions

Four publications A2 to A5 are included in this section:

- In A2, we explored the spatially differentiated trade-offs between economic benefits and roadside environmental impacts at a continental scale, using the model of Spatial Stratified Heterogeneity (SSH).

Related publication: (A2) Luo, P., Song, Y. and Wu, P., 2021. Spatial disparities in trade-offs: economic and environmental impacts of road infrastructure on continental level. GIScience and Remote Sensing, 58(5), pp.756-775.

The study uncovers substantial spatial disparities in the effects of road infrastructure on the economy and the roadside environment. In major cities like Sydney and Melbourne, economic growth exacerbates environmental pressure, whereas in suburban and rural areas, the roadside environment has improved.

- In A3, we developed the model of Geographically Optimal Zones-based Heterogeneity model (GOZH) to achieve a more accurate representation of spatial correlations.

Related publication: (A3) Luo, P., Song, Y., Huang, X., Ma, H., Liu, J., Yao, Y. and Meng, L., 2022. Identifying determinants of spatio-temporal disparities in soil moisture of the Northern Hemisphere using a geographically optimal zones-based heterogeneity model. ISPRS Journal of Photogrammetry and Remote Sensing, 185, pp.111-128

In the SSH model, spatial discretization is typically performed using equal intervals, quantiles, or geometric divisions, without any optimization in the process. The outcomes of this method are influenced by the rules for spatial discretization. Consequently, the Power of Determinants (PD) cannot fully elucidate the spatial association between explanatory and response variables. Studies have indicated significant underestimation of PD in the classic SSH model (Song et al. 2020; Luo et al. 2022). Moreover, the process of spatial discretization in spatial stratified heterogeneity should be consistent, whether for a single explanatory variable or multiple variables, a challenge not yet addressed by current algorithms. Therefore, we discard the distinction between the effects of single factors and interactions in the current model. We argue that the most accurate representation of spatial analysis heterogeneity for a given set of explanatory variables, be it single or multiple, is the one that yields the highest PD value among countless spatial discretization methods. We define this value as Ω , which can also be referred as the Optimal PD (OPD).

$$\Omega = \max(\gamma) = 1 - \frac{\min(SSW_{X,D})}{SST} \quad (1)$$

where X represents the explanatory variable, and D denotes the spatially discretized variable. Therefore, within Ω , the Sum of Squares Within (SSW) is a function of X and D.

To solve for the Ω value, we transform the problem of determining spatially stratified heterogeneity into an optimization problem and incorporate machine learning algorithms, utilizing a stepwise optimization approach (see Figure 1).

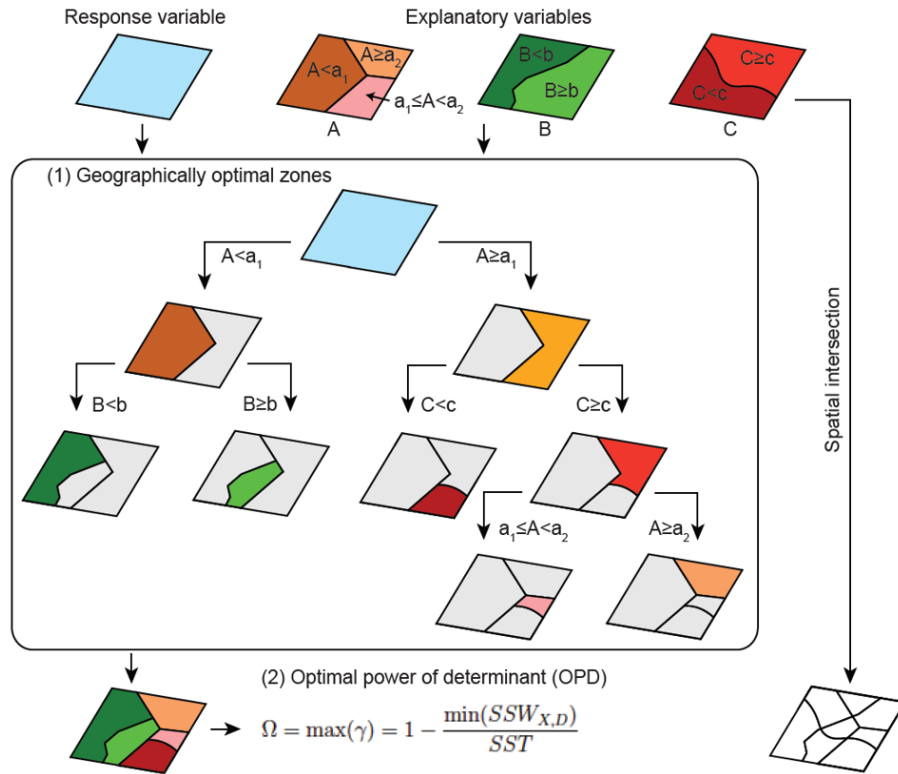


Figure 3-5. The illustration of the GOZH model

- In A4, we attempted to enhance the interpretability of our model. Drawing on game theory, we analyzed the collinearity of each explanatory variable in their interactions.

Related publication: (A4) Li, Y.[#], Luo, P.[#](co-first), Song, Y., Zhang, L., Qu, Y and Hou, Z., 2023. A locally explained heterogeneity model for examining wetland disparity. International Journal of Digital Earth, 13(2), p.4533-4552.

The objective of this study was to demystify the "black box" and ascertain the contribution of individual variables to the OPD. We introduced the Locally Explained Heterogeneity Model(LESH), which, in conjunction with the SHAP (Shapley Additive exPlanations) and SSH (Spatially Stratified Heterogeneity) models, allows for a comprehensive explanation of each variable's contribution, irrespective of the number of variables or the complexity of their interactions. We decomposed the PD to ascertain the contribution of each variable, employing the Shapley value for this purpose. The Shapley value, a concept from cooperative game theory, measures the contribution of participants to the collective payoff of a cooperative game. It is a method for distributing the total payoff of the cooperative game among the participants, to fairly assess each participant's contribution. In our study, we leveraged the concept of the Shapley value to calculate the contribution of each explanatory variable to the PD, denoting the calculated values as the Shapley Power of Determinants (SPD).

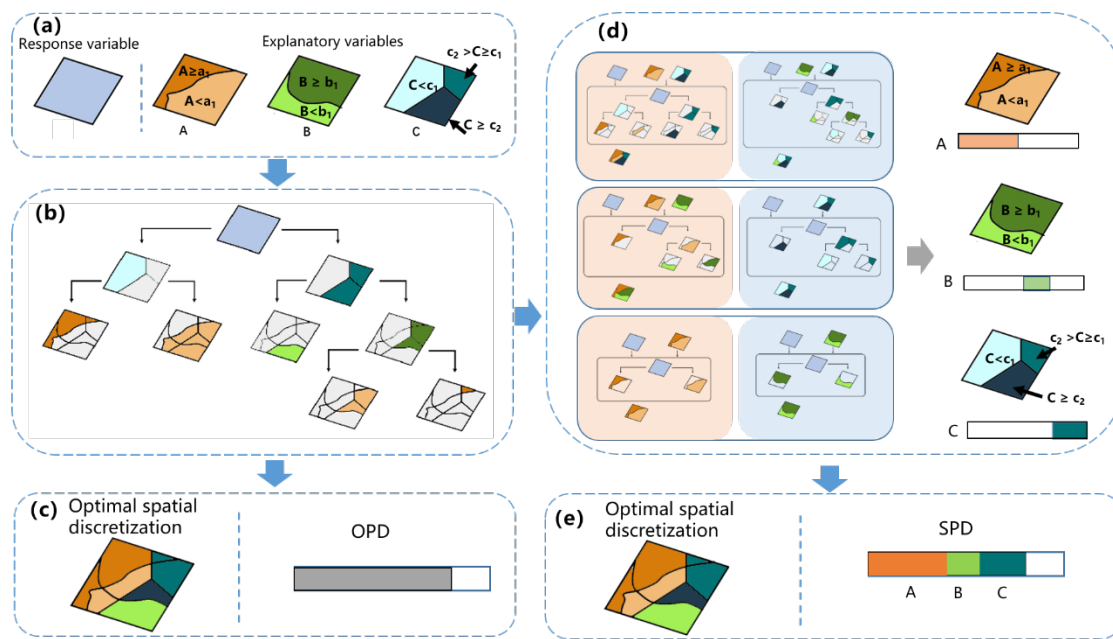


Figure 3-6. The computational workflow of the LESH model. The steps (a-c) represent the process of calculating OPD, while steps (a, b, d, e) depict the process of calculating the SPD.

- In A5, we developed a Geographical Pattern Interaction model capable of exploring spatial variability parameters.

Related publication:(A5) Luo, P., et al. (2023). Measuring univariate effects in the interaction of geographical patterns. International Journal of Geographical Information Science. (Under Review)

Building on the LESH model, we endeavored to solve the correlations between variables at each location. Under the theoretical assumption that similarity in spatial patterns can indicate spatial associations, we developed global and local indicators of spatial patterns. These indicators are based on the interactions of spatial patterns of different variables, and we introduced the Shapley value to identify the contributions of different variables at both global and local levels.

This study introduces a Geographical Pattern Interaction (GPI) model to analyze univariate effects within the context of pattern interaction among variables. The GPI model consists of three main components (Figure 3-7):

1. Generation of GPI for the response variable based on the spatial patterns of multiple explanatory variables, facilitated by the Geographically Optimal Zones-based Heterogeneity (GOZH) model. This aims to identify optimal geographic zones for variable combinations and establish the best geographical partition for all explanatory variables.

2. Computation of global univariate effects in GPI, where the variance for each geographic partition under various variable combinations is calculated to assess the overall relationship between explanatory and response variables. This includes interactions between single and multiple explanatory variables, using spatially stratified heterogeneity and the SHAP interpretable machine learning algorithm to quantify the contribution of individual variables in these interactions.

3. Assessment of local univariate effects in GPI, focusing on the local effects of GPI, local univariate effects, and characteristics like nonlinearity, dominant local variables, and bivariate effects. The model calculates mean responses for each geographic partition under different variable combinations and employs SHAP to determine the variable contributions to the geographic classification, reflecting the relationship between explanatory and response variables in each region.

The GPI model was applied to identify factors affecting the homeless rate in Australia, demonstrating its practical utility in real-world applications.

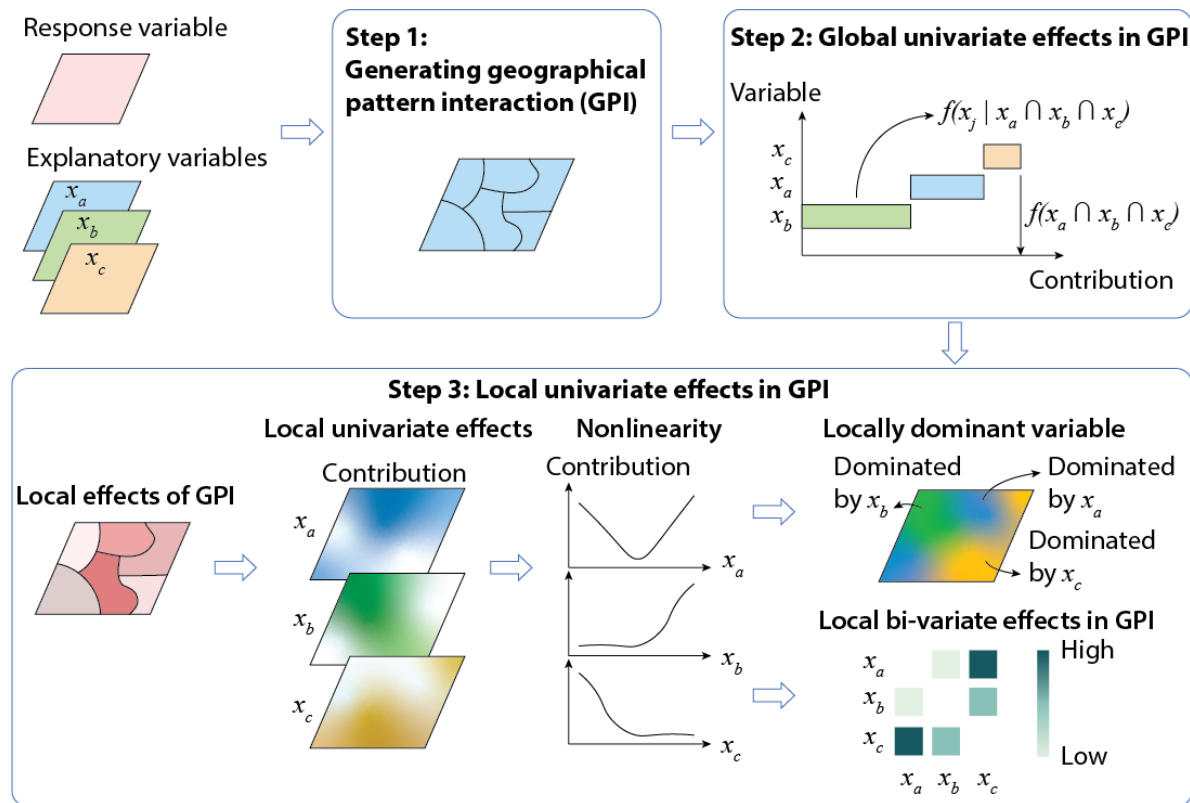


Figure 3-7. The framework of GPI model.

3.3 Spatial prediction with biased and sparse data

To address the challenges of scarcity and bias in spatial data for spatial prediction, we have selected two cases for exploration: spatial interpolation (A6) and spatial extrapolation prediction (A7).

- In A6, we developed a Generalized Heterogeneity Model (GHM) for large scale spatial interpolation

Related publication: (A6) Luo, P., Song, Y., Zhu, D., Cheng J. and Meng L. A Generalized Spatial Heterogeneity Model for Interpolation. 2022 International Journal of Geographical Information Science, 37(3), 634-659

Spatial heterogeneity often manifests as geographic variables existing in multiple homogeneous spatial strata. Therefore, a straightforward and effective approach to apply geostatistical models to spatially second-order non-stationary surfaces, is to divide the geographical space into several strata with relatively strong homogeneity. These subregions often satisfy the assumption of spatial second-order stationarity, allowing for spatial interpolation within each. This method is known as Stratified Kriging (StK). However, StK has two main drawbacks: firstly, the interpolation within each stratum ignores the numerical information of other strata, leading to a loss in accuracy. Secondly, the spatial division process and subsequent independent interpolation in each stratum can result in unrealistic abrupt changes and discontinuities along the strata boundaries, often contradicting our understanding of geospatial continuity.

The motivation for this research is to achieve accurate and reliable spatial predictions for large-scale geographical environments, considering both the existence of spatial strata and the spatial dependencies along the strata boundaries. A practical solution is to use information from other strata when predicting a particular stratum, ensuring overall accuracy while making reasonable estimations at the strata boundaries. For example, when interpolating population density in urban-rural transition zones, data from both urban and rural areas provide necessary information. In interpolating elevation in plain-plateau transition zones, it is essential to consider the mixed characteristics of plateau and plain elevations.

Our proposed solution employs Area-To-Area Kriging (ATAK) to describe long-distance spatial relationships, widely used for modeling spatial data at different scales (Figure 3-8). We believe that ATAK can calculate the weights of other strata, representing the spatial associations between different strata. Based on our fundamental assumption that the scale of spatial relationships of geographic variables is distance-dependent, we differentiate spatial relationships in geostatistical modeling into neighborhood and long-distance spatial relationships. Neighborhood spatial relationships are described using pairwise points and

solved using ordinary kriging. In contrast, long-distance global spatial relationships, representing larger scales or areas, are solved using ATAK. Using this method, our proposed generalized heterogeneity model extends geostatistical modeling spatially, allowing for more accurate large-scale spatial predictions under the premise of spatial second-order non-stationarity by separately modeling local and global spatial relationships.

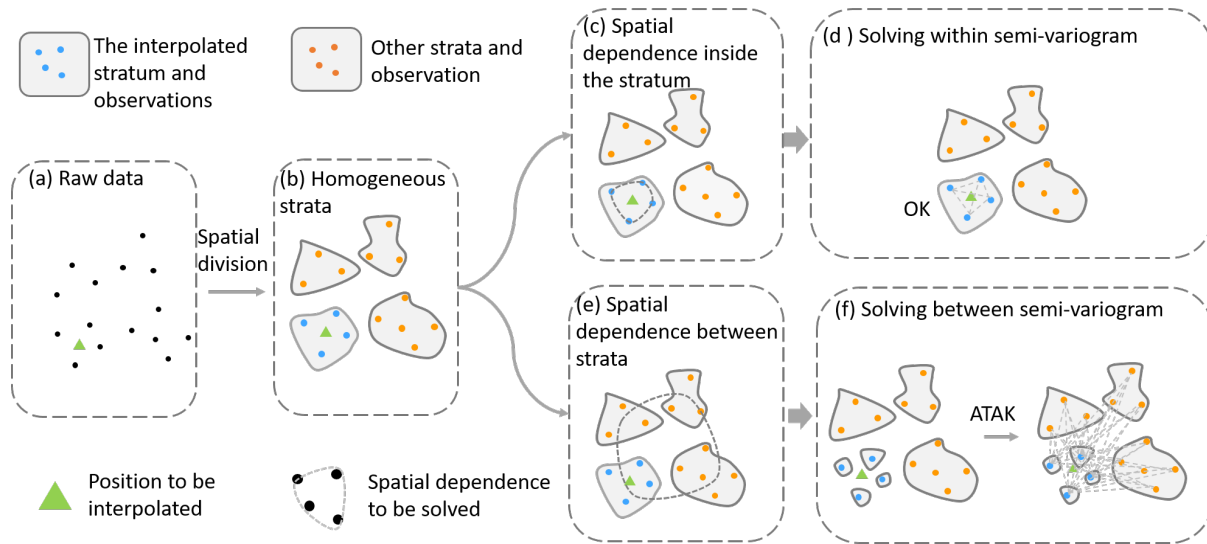


Figure 3-8. The framework of GHM.

- In A7, to address the issue of misleading information caused by the inherent bias of spatial data in spatial predictions, we introduced the concept of anomaly detection.

Related publication: (A7) Yao, Y., Dong, A., Liu, Z., Jiang, Y., Guo, Z., Cheng, J., Guan, Q. and Luo, P. (correspondence, project lead), 2023. Extracting the pickpocketing information implied in the built environment by treating it as the anomalies. Cities, 143, p.104575.*

Predicting certain geospatial phenomena, such as crime or traffic accidents, often involves biased and sparse data, leading to model underfitting, significant bias, and even potential ethical issues. In response to this problem, we propose treating geospatial phenomena with biased and sparse samples as "anomalies." This allows us to use the abundantly available "normal samples" to build models and implement large-scale spatial predictions through anomaly detection.

In A8, we demonstrated this approach using pickpocketing risk prediction task, where, with only limited crime location points and based on street view images, we achieved fine-grained risk prediction across the entire city of Shenzhen, China. We collected 154,868 street view images from Shenzhen with the aim of

predicting the latent crime risk in each image (Figure 3-9). However, in 2018, Shenzhen City only recorded the locations of 682 pickpocketing crime. This presents a typical sparse data spatial prediction problem. We interpret the prediction of crime risk from street view images as anomaly detection. Despite having only 682 spatial locations of crimes, resulting in a limited number of street view images, we can extract features from images that are not close to crime locations, which we term as normal features. We then compute the features for each image in the test set using the same model and calculate its similarity to the normal features. The lower the similarity, the higher the underlying crime risk in the image.

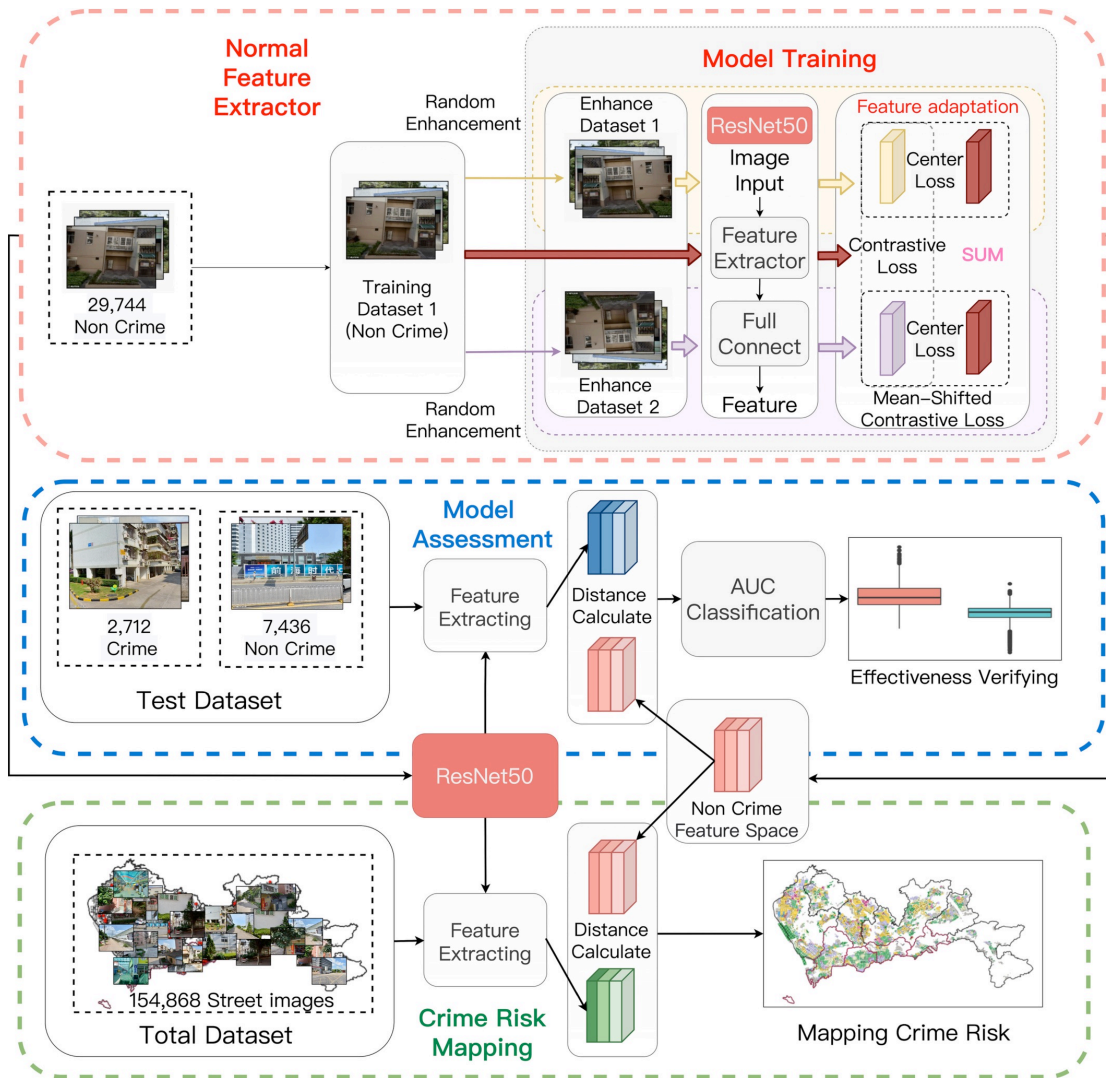


Figure 3-9. The framework of estimating crime risk based on the anomaly detection

4. Conclusion and outlook

4.1 Conclusion

The distinction between spatial tasks and other types of tasks is attributed to the inherent characteristics of spatial data that differ from those of other data types. The uniqueness of spatial data is primarily described through spatial effects and spatial relationships. Currently, many spatial relationships are not yet adequately modeled, which limits the performance of spatial analysis.

This thesis aims to modeling the generalized spatial association and addressed research gaps in three spatial analysis tasks – description, explanation and prediction. We developed new spatial models based on a deeper understanding of spatial data. The research questions have been explored with the following findings:

- Overlapping community structures can be uncovered through graph generation tasks. We proposed a mechanism model for generating spatial interactions based on overlapping community structures and transformed the task of inferring community structures from spatial interaction data into a graph generation task.
- Non-linear relationships and interactions among geographic variables can be mined from spatial data with weak assumptions, by analyzing the similarity and interactions of spatial patterns.
- When using biased and sparse spatial data for spatial prediction:
 - (Interpolation) Large-scale unbiased estimation can be achieved by constructing spatial relationships at different scales to address the modeling issues in non-stationary spatial areas;
 - (Extrapolation) The problem of misleading conclusions caused by statistical and spatial biases in data can be addressed using the concept of anomaly detection.

4.2 Limitations and outlook

Space possesses its uniqueness, and there is a broad spectrum of research efforts, including this thesis, aimed at mitigating the negative impacts of this uniqueness on spatial analysis. These efforts focus on harnessing appropriate modeling methods to fully exploit the rich information contained within spatial data. Nonetheless, understanding and exploring the characteristics of geospatial data remains a challenging process that is far from complete.

For descriptive tasks, this thesis concentrates on spatial interaction data and community structure. However, the connotation of spatial patterns extends beyond this scope. Besides spatial interaction data that can be

represented as a network model, certain spatial data need to be characterized as field models and object models. Pattern recognition for these types of spatial data is not covered in this thesis.

In terms of explanatory tasks, the methods uncovered in this thesis effectively analyze the spatial correlations between different geographical variables. Yet, the proposed models have not addressed spatial causality, which is a valuable direction for future research. Additionally, the method proposed in this thesis aims to analyze spatial correlations by examining similarities in spatial patterns. We attempt to extract this correlation purely through numerical correlations. Therefore, we do not explicitly consider geographic location information, such as incorporating latitude and longitude as features or parameters into the model. Due to the collinearity between location features and many geographic variables, explicitly considering them may actually lead to inaccurate results. However, some scholars argue that spatial analysis requires explicit consideration of spatial location. We do not fully agree with this view but acknowledges that in further research, it is worth exploring the impact of spatial location when analyzing correlations between variables. This is because in some cases, the spatial distribution of geographic variables may not primarily result from the influence of another variable but from their own spatial processes, such as spatial diffusion. Therefore, eliminating this aspect of influence when analyzing the correlation between different variables is an important topic.

For predictive tasks, the first work involves large-scale spatial interpolation, necessitating the reasonable partitioning of space into smaller, homogeneous regions. In this step, a simplistic method was employed, dividing the study area by latitude and longitude, which may not align with the spatial distribution characteristics of geographical variables. In future work, more precise spatial partitioning methods should be employed, especially those that consider the spatial distribution patterns of geographical variables thoroughly. In extending the spatial scope of this work, street view images were used to predict crime risk. The hypothesis was that the distribution of some geographical phenomena in space could be considered anomalous. Therefore, only samples where these phenomena had not occurred were used for model training. However, in many cases, it is challenging to ensure that locations in the training data without recorded crimes are genuinely free of crime risk, possibly because crimes in these areas went unrecorded or undetected by the police. This uncertainty is a characteristic of spatial data, limited by the costs and time constraints of sampling, meaning that spatial data always contain uncertainties regarding the actual distribution of geographical variables and phenomena. Future research on spatial prediction tasks based on anomaly detection should incorporate as much prior knowledge as possible and leverage expert experience to select training samples, thus maximizing the reliability of the training data.

Additionally, most data utilized in this thesis are sourced from open-access databases and span multiple regions (e.g., USA, China, Australia). These datasets are largely reproducible and reusable. All remote sensing data employed in the thesis are collected from the Google Earth Engine (GEE). For instance, the NDVI data in A2, precipitation data in A3, and DEM data in A4. Furthermore, a significant portion of the data sources are derived from official census data, such as the Australian homeless rate data in A5 sourced from the Australian Bureau of Statistics, and the pickpocketing data in Shenzhen in A8 sourced from openly available court judgment data in China. However, there still exist certain non-open-source data in this thesis, such as the device-level travel trajectory data in A1, which is obtained from a commercial entity (i.e., PlaceIQ³). This presents a limitation of this thesis. Although we believe the contribution of A1 primarily lies in its methodological approach, validating models using open-source data and publishing results in the future remains a worthwhile direction.

Finally, throughout the process of conducting a series of studies, i.e., exploring more comprehensive and accurate spatial relationship modeling methods, we have been repeatedly inspired by visual analytics. For example, in the second part of the thesis, while exploring the spatial relationships of different geographical variables, the inspiration stemmed from the intuitive understanding that humans have regarding the correlation between two variables: correlated variables exhibit similar spatial distributions. This aligns with the realm of geovisual analytics. This thesis endeavors to translate conclusions from visual analytics into mathematical language and construct models to explore the correlation of geographical variables. This thesis has shown the feasibility of utilizing visual analytics to investigate spatial relationships. We believe this is partly attributed to human visual intuition in comprehending the geographical world. In future work, we anticipate that integrating advanced multivariate visualization methods will further extend the model's applicability and enhance our insight and understanding of the intricate relationships within geographical data.

Bibliography

- Anselin, Luc. 1988. *Spatial Econometrics: Methods and Models*. Vol. 4. Studies in Operational Regional Science. Dordrecht: Springer Netherlands. <https://doi.org/10.1007/978-94-015-7799-1>.
- . 1992. 'Spatial Data Analysis with GIS: An Introduction to Application in the Social Sciences (92-10)', August. <https://escholarship.org/uc/item/58w157nm>.
- . 1995. 'Local Indicators of Spatial Association—LISA'. *Geographical Analysis* 27 (2): 93–115. <https://doi.org/10.1111/j.1538-4632.1995.tb00338.x>.
- . 2010. 'Thirty Years of Spatial Econometrics: Thirty Years of Spatial Econometrics'. *Papers in Regional Science* 89 (1): 3–25. <https://doi.org/10.1111/j.1435-5957.2010.00279.x>.

³ <https://www.precisely.com/about-us/placeiq-is-now-part-of-precisely>

- Anselin, Luc, and Pedro Amaral. 2023. 'Endogenous Spatial Regimes'. *Journal of Geographical Systems*, June. <https://doi.org/10.1007/s10109-023-00411-2>.
- A.S., Hedayat, and Sinha B.K. 1991. *Design and Inference in Finite Population Sampling*. Wiley. <https://www.wiley.com/en-cn/Design+and+Inference+in+Finite+Population+Sampling-p-9780471880738>.
- Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. 'Fast Unfolding of Communities in Large Networks'. *Journal of Statistical Mechanics: Theory and Experiment* 2008 (10): P10008. <https://doi.org/10.1088/1742-5468/2008/10/P10008>.
- Capello, Roberta. 2009. 'Spatial Spillovers and Regional Growth: A Cognitive Approach'. *European Planning Studies* 17 (5): 639–58. <https://doi.org/10.1080/09654310902778045>.
- Chang, Victor. 2021. 'An Ethical Framework for Big Data and Smart Cities'. *Technological Forecasting and Social Change* 165 (April): 120559. <https://doi.org/10.1016/j.techfore.2020.120559>.
- Chen, Yanguang. 2015. 'The Distance-Decay Function of Geographical Gravity Model: Power Law or Exponential Law?' *Chaos, Solitons & Fractals* 77 (August): 174–89. <https://doi.org/10.1016/j.chaos.2015.05.022>.
- Chin, Wei-Chien-Benny, Tzai-Hung Wen, Clive E. Sabel, and I.-Hsiang Wang. 2017. 'A Geo-Computational Algorithm for Exploring the Structure of Diffusion Progression in Time and Space'. *Scientific Reports* 7 (1): 12565. <https://doi.org/10.1038/s41598-017-12852-z>.
- Chou, Yue -Hong. 1995. 'Spatial Pattern and Spatial Autocorrelation'. In *Spatial Information Theory A Theoretical Basis for GIS*, edited by Andrew U. Frank and Werner Kuhn, 988:365–76. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-60392-1_24.
- Cressie, Noel. 1990. 'The Origins of Kriging'. *Mathematical Geology* 22: 239–52.
- De Marsily, Gh., F. Delay, J. Gonçalves, Ph. Renard, V. Teles, and S. Violette. 2005. 'Dealing with Spatial Heterogeneity'. *Hydrogeology Journal* 13 (1): 161–83. <https://doi.org/10.1007/s10040-004-0432-3>.
- Dong, Lei, Fabio Duarte, Gilles Duranton, Paolo Santi, Marc Barthelemy, Michael Batty, Luís Bettencourt, et al. 2024. 'Defining a City — Delineating Urban Areas Using Cell-Phone Data'. *Nature Cities*, January, 1–9. <https://doi.org/10.1038/s44284-023-00019-z>.
- Fotheringham, A. Stewart. 1981. 'Spatial Structure and Distance-Decay Parameters'. *Annals of the Association of American Geographers* 71 (3): 425–36. <https://doi.org/10.1111/j.1467-8306.1981.tb01367.x>.
- Fotheringham, A. Stewart, Chris Brunson, and Martin Charlton. 2003. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. John Wiley & Sons.
- Fotheringham, A. Stewart, and Ziqi Li. 2023. 'Measuring the Unmeasurable: Models of Geographical Context'. *Annals of the American Association of Geographers* 113 (10): 2269–86. <https://doi.org/10.1080/24694452.2023.2227690>.
- Fotheringham, A. Stewart, and P.A. Rogerson. 2008. *The SAGE Handbook of Spatial Analysis*. SAGE. <https://uk.sagepub.com/en-gb/eur/the-sage-handbook-of-spatial-analysis/book227940>.
- Gao, Bingbo, Maogui Hu, Jinfeng Wang, Chengdong Xu, Ziyue Chen, Haimei Fan, and Haiyuan Ding. 2020. 'Spatial Interpolation of Marine Environment Data Using P-MSN'. *International Journal of Geographical Information Science* 34 (3): 577–603. <https://doi.org/10.1080/13658816.2019.1683183>.
- Gao, Song, Yu Liu, Yaoli Wang, and Xiujun Ma. 2013. 'Discovering Spatial Interaction Communities from Mobile Phone Data'. *Transactions in GIS* 17 (3): 463–81. <https://doi.org/10.1111/tgis.12042>.
- Gelfand, Alan E., Hyon-Jung Kim, C. F. Sirmans, and Sudipto Banerjee. 2003. 'Spatial Modeling with Spatially Varying Coefficient Processes'. *Journal of the American Statistical Association* 98 (462): 387–96.
- Getis, Arthur, and J. K. Ord. 1992. 'The Analysis of Spatial Association by Use of Distance Statistics'. *Geographical Analysis* 24 (3): 189–206. <https://doi.org/10.1111/j.1538-4632.1992.tb00261.x>.

- Goodchild, M F, L Anselin, and U Deichmann. 1993. 'A Framework for the Areal Interpolation of Socioeconomic Data'. *Environment and Planning A: Economy and Space* 25 (3): 383–97. <https://doi.org/10.1068/a250383>.
- Goodchild, Michael F. 1989. 'Modeling Error in Objects and Fields'. In *The Accuracy Of Spatial Databases*. CRC Press.
- . 1992. 'Geographical Data Modeling'. *Computers & Geosciences, GIS Design Models*, 18 (4): 401–8. [https://doi.org/10.1016/0098-3004\(92\)90069-4](https://doi.org/10.1016/0098-3004(92)90069-4).
- Griffith, Daniel A. 2003. 'Spatial Filtering'. In *Spatial Autocorrelation and Spatial Filtering: Gaining Understanding Through Theory and Scientific Visualization*, edited by Daniel A. Griffith, 91–130. Advances in Spatial Science. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-24806-4_4.
- . 2005. 'Effective Geographic Sample Size in the Presence of Spatial Autocorrelation'. *Annals of the Association of American Geographers* 95 (4): 740–60.
- Grover, Aditya, and Jure Leskovec. 2016. 'Node2vec: Scalable Feature Learning for Networks'. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 855–64. KDD '16. New York, NY, USA: Association for Computing Machinery. <https://doi.org/10.1145/2939672.2939754>.
- Guo, Diansheng, Hai Jin, Peng Gao, and Xi Zhu. 2018. 'Detecting Spatial Community Structure in Movements'. *International Journal of Geographical Information Science* 32 (7): 1326–47. <https://doi.org/10.1080/13658816.2018.1434889>.
- Guo, Jiangang, Jinfeng Wang, Chengdong Xu, and Yongze Song. 2022. 'Modeling of Spatial Stratified Heterogeneity'. *GIScience & Remote Sensing* 59 (1): 1660–77. <https://doi.org/10.1080/15481603.2022.2126375>.
- Haining, Robert. 1990. *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511623356>.
- Hong, Ye, and Yao Yao. 2019. 'Hierarchical Community Detection and Functional Area Identification with OSM Roads and Complex Graph Theory'. *International Journal of Geographical Information Science* 33 (8): 1569–87. <https://doi.org/10.1080/13658816.2019.1584806>.
- Jia, Tao, Chenxi Cai, Xin Li, Xi Luo, Yuanyu Zhang, and Xuesong Yu. 2022. 'Dynamical Community Detection and Spatiotemporal Analysis in Multilayer Spatial Interaction Networks Using Trajectory Data'. *International Journal of Geographical Information Science* 36 (9): 1719–40. <https://doi.org/10.1080/13658816.2022.2055037>.
- Kipf, Thomas N., and Max Welling. 2017. 'Semi-Supervised Classification with Graph Convolutional Networks'. arXiv. <https://doi.org/10.48550/arXiv.1609.02907>.
- Kwan, Mei-Po. 2012. 'The Uncertain Geographic Context Problem'. *Annals of the Association of American Geographers* 102 (5): 958–68. <https://doi.org/10.1080/00045608.2012.687349>.
- Lam, Nina Siu-Ngan. 1983. 'Spatial Interpolation Methods: A Review'. *The American Cartographer* 10 (2): 129–50. <https://doi.org/10.1559/152304083783914958>.
- Li, Mingxiao, Song Gao, Feng Lu, Kang Liu, Hengcai Zhang, and Wei Tu. 2021. 'Prediction of Human Activity Intensity Using the Interactions in Physical and Social Spaces through Graph Convolutional Networks'. *International Journal of Geographical Information Science* 35 (12): 2489–2516. <https://doi.org/10.1080/13658816.2021.1912347>.
- Li, Ziqi. 2022. 'Extracting Spatial Effects from Machine Learning Model Using Local Interpretation Method: An Example of SHAP and XGBoost'. *Computers, Environment and Urban Systems* 96 (September): 101845. <https://doi.org/10.1016/j.compenvurbsys.2022.101845>.
- Liu, Yu, Keli Wang, Xiaoyue Xing, Hao Guo, Weiyu Zhang, Qingyao Luo, Song Gao, et al. 2023. 'On Spatial Effects in Geographical Analysis'. *Acta Geographica Sinica* 78 (3): 517. <https://doi.org/10.11821/dlxb202303001>.
- Liu, Yu, Xin YAO, Yongxi GONG, Chaogui KANG, Xun SHI, Fahui WANG, Jiao'e WANG, et al. 2020. 'Analytical methods and applications of spatial interactions in the era of big data'. *Acta Geographica Sinica* 75 (7): 1523–38. <https://doi.org/10.11821/dlxb202007014>.

- Longley, Paul A., Michael F. Goodchild, David J. Maguire, and David W. Rhind. 2015. *Geographic Information Science and Systems*. John Wiley & Sons.
- Luo, Peng, and Yongze Song. 2021. 'A Spatial Second-Order Non-Stationary Interpolation Method for Large Area Mapping'. *Abstracts of the ICA 3* (December): 1–1. <https://doi.org/10.5194/ica-abs-3-187-2021>.
- Luo, Peng, Yongze Song, Xin Huang, Hongliang Ma, Jin Liu, Yao Yao, and Liqiu Meng. 2022. 'Identifying Determinants of Spatio-Temporal Disparities in Soil Moisture of the Northern Hemisphere Using a Geographically Optimal Zones-Based Heterogeneity Model'. *ISPRS Journal of Photogrammetry and Remote Sensing* 185 (March): 111–28. <https://doi.org/10.1016/j.isprsjprs.2022.01.009>.
- Luo, Peng, Yongze Song, Di Zhu, Junyi Cheng, and Liqiu Meng. 2023. 'A Generalized Heterogeneity Model for Spatial Interpolation'. *International Journal of Geographical Information Science* 37 (3): 634–59. <https://doi.org/10.1080/13658816.2022.2147530>.
- Luo, Peng, and Di Zhu. 2022. 'Sensing Overlapping Geospatial Communities from Human Movements Using Graph Affiliation Generation Models'. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, 1–9. Seattle Washington: ACM. <https://doi.org/10.1145/3557918.3565862>.
- . 2024. 'Uncover the Nature of Overlapping Community in Cities'. arXiv. <https://doi.org/10.48550/arXiv.2402.00222>.
- Mitas, L., and H. Mitasova. n.d. 'Spatial Interpolation'. In *Geographical Information Systems: Principles, Techniques, Management and Applications*, 1:481–92. New York: NY: Wiley. Accessed 7 March 2024. http://fatra.cnr.ncsu.edu/~hmitaso/gmslab/papers/mitas_mitasova_1999_2005.pdf.
- Moran, P. A. P. 1950. 'Notes on Continuous Stochastic Phenomena'. *Biometrika* 37 (1/2): 17–23. <https://doi.org/10.2307/2332142>.
- Newman, M. E. J. 2013. 'Community Detection and Graph Partitioning'. *EPL (Europhysics Letters)* 103 (2): 28003. <https://doi.org/10.1209/0295-5075/103/28003>.
- O'Kelly, M. E. 2009. 'Spatial Interaction Models'. In *International Encyclopedia of Human Geography*, edited by Rob Kitchin and Nigel Thrift, 365–68. Oxford: Elsevier. <https://doi.org/10.1016/B978-008044910-4.00529-0>.
- OLIVER, M. A., and R. WEBSTER. 1990. 'Kriging: A Method of Interpolation for Geographical Information Systems'. *International Journal of Geographical Information Systems* 4 (3): 313–32. <https://doi.org/10.1080/02693799008941549>.
- S. Granovetter, Mark. 1973. 'The Strength of Weak Ties.Pdf'. *American Journal of Sociology* 78 (6): 1360–80.
- Sachdeva, Mehak, A. Stewart Fotheringham, Ziqi Li, and Hanchen Yu. 2022. 'Are We Modelling Spatially Varying Processes or Non-Linear Relationships?' *Geographical Analysis* 54 (4): 715–38. <https://doi.org/10.1111/gean.12297>.
- Setayesh, Amin, Zhivar Sourati Hassan Zadeh, and Behnam Bahrak. 2022. 'Analysis of the Global Trade Network Using Exponential Random Graph Models'. *Applied Network Science* 7 (1): 1–19. <https://doi.org/10.1007/s41109-022-00479-7>.
- Song, Yongze, Jinfeng Wang, Yong Ge, and Chengdong Xu. 2020. 'An Optimal Parameters-Based Geographical Detector Model Enhances Geographic Characteristics of Explanatory Variables for Spatial Heterogeneity Analysis: Cases with Different Types of Spatial Data'. *GIScience & Remote Sensing* 57 (5): 593–610. <https://doi.org/10.1080/15481603.2020.1760434>.
- Tiefelsdorf, Michael, and Daniel A Griffith. 2007. 'Semiparametric Filtering of Spatial Autocorrelation: The Eigenvector Approach'. *Environment and Planning A: Economy and Space* 39 (5): 1193–1221. <https://doi.org/10.1068/a37378>.
- Wang, Jin-Feng, A. Stein, Bin-Bo Gao, and Yong Ge. 2012. 'A Review of Spatial Sampling'. *Spatial Statistics* 2 (December): 1–14. <https://doi.org/10.1016/j.spasta.2012.08.001>.

- Wang, Yujing, Yi Deng, Fu Ren, Ruoxin Zhu, Pei Wang, Tian Du, and Qingyun Du. 2020. 'Analysing the Spatial Configuration of Urban Bus Networks Based on the Geospatial Network Analysis Method'. *Cities* 96 (January): 102406. <https://doi.org/10.1016/j.cities.2019.102406>.
- Yao, Yao, Chenqi Feng, Jiteng Xie, Xiaoqin Yan, Qingfeng Guan, Jian Han, Jiaqi Zhang, Shuliang Ren, Yuyun Liang, and Peng Luo. 2023. 'A Site Selection Framework for Urban Power Substation at Micro-scale Using Spatial Optimization Strategy and Geospatial Big Data'. *Transactions in GIS*, August, tgis.13093. <https://doi.org/10.1111/tgis.13093>.
- Yao, Yao, Jiale Wang, Ye Hong, Chen Qian, Qingfeng Guan, Xun Liang, Liangyang Dai, and Jinbao Zhang. 2021. 'Discovering the Homogeneous Geographic Domain of Human Perceptions from Street View Images'. *Landscape and Urban Planning* 212 (August): 104125. <https://doi.org/10.1016/j.landurbplan.2021.104125>.
- Zhao, Xuehua, Bo Yang, Xueyan Liu, and Huiling Chen. 2017. 'Statistical Inference for Community Detection in Signed Networks'. *Physical Review E* 95 (4): 042313. <https://doi.org/10.1103/PhysRevE.95.042313>.
- Zhou, Zhixin, and Arash A. Amini. 2018. 'Analysis of Spectral Clustering Algorithms for Community Detection: The General Bipartite Setting'. arXiv. <https://doi.org/10.48550/arXiv.1803.04547>.
- Zhu, Di, and Guofeng Cao. 2023. 'Intelligent Spatial Prediction and Interpolation Methods'. In *Handbook of Geospatial Artificial Intelligence*. CRC Press.
- Zhu, Di, Yu Liu, Xin Yao, and Manfred M. Fischer. 2022. 'Spatial Regression Graph Convolutional Neural Networks: A Deep Learning Paradigm for Spatial Multivariate Distributions'. *GeoInformatica* 26 (4): 645–76. <https://doi.org/10.1007/s10707-021-00454-x>.

Appendix: Publications

A1. Sensing overlapping geospatial communities from human movements using graph affiliation generation models

Reference: Luo, P. and Zhu, D., 2022, November. Sensing overlapping geospatial communities from human movements using graph affiliation generation models. In Proceedings of the 5th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery (pp. 1-9).

Sensing overlapping geospatial communities from human movements using graph affiliation generation models

Peng Luo
peng.luo@tum.de
Technical University of Munich
Munich, Bavaria, Germany

Di Zhu*
dizhu@umn.edu
University of Minnesota
Minneapolis, USA

ABSTRACT

Geographical units densely connected by human movements can be treated as a geospatial community. Detecting geospatial communities in a mobility network reveals key characteristics of human movements and urban structures. Recent studies have found communities can be overlapping in that one location may belong to multiple communities, posing great challenges to classic disjoint community detection methods that only identify single-affiliation relationships. In this work, we propose a Geospatial Overlapping Community Detection (GOCD) framework based on graph generation models and graph-based deep learning. GOCD aims to detect geographically overlapped communities regarding the multiplex connections underlying human movements, including weak and long-range ties. The detection process is formalized as deriving the optimized probability distribution of geographic units' community affiliations in order to generate the spatial network, i.e., the most reasonable community affiliation matrix given the observed network structure. Further, a graph convolutional network (GCN) is introduced to approach the affiliation probabilities via a deep learning strategy. The GOCD framework outperformed existing baselines on non-spatial benchmark datasets in terms of accuracy and speed. A case study of mobile positioning data in the Twin Cities Metropolitan Area (TCMA), Minnesota, was presented to validate our model on real-world human mobility networks. Our empirical results unveiled the overlapping spatial structures of communities, the overlapping intensity for each CBG, and the spatial heterogeneous structure of community affiliations in the Twin Cities.

CCS CONCEPTS

• **Networks** → **Topology analysis and generation; Network mobility**; • **Human-centered computing** → **Ubiquitous and mobile devices**.

*Contact author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

GeoAI '22, November 1, 2022, Seattle, WA, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9532-8/22/11...\$15.00
<https://doi.org/10.1145/3557918.3565862>

KEYWORDS

Community detection, Graph convolutional networks, Overlapping, Human mobility, Urban structure

ACM Reference Format:

Peng Luo and Di Zhu. 2022. Sensing overlapping geospatial communities from human movements using graph affiliation generation models. In *The 5th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery (GeoAI '22) (GeoAI '22), November 1, 2022, Seattle, WA, USA*. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3557918.3565862>

1 INTRODUCTION

Geographic units with strong human movement connections can be called geospatial communities [11]. Geographic units within the same geospatial community are more closely connected than those belonging to different communities [6, 8]. Social services, such as police deployment and medical services, are often optimized as community-based resource allocation. Policymakers need to consider reasonable policies based on the spatial distribution of communities. Research has shown that geographic space can be well characterized by graph structures, where geographic units are formalized as the nodes, and the connections between geographic units are the edges. Therefore, the sensing of geospatial communities, i.e., assigning a community label for each location in the network, can be seen as a community detection task in graph-structured spatial networks [10]. Detecting communities in spatial networks is essential in understanding human activities, socioeconomics, urban structure, etc. through looking at the socio-economic interactions [22], and even the hierarchical relationship between regions [10].

Traditional community detection methods mainly focus on identifying disjoint communities (as shown in Figure 1 b), which means that each node may belong to only one community [1, 3]. While in the real world, communities could overlap (see Figure 1 c) [5, 13–15]. For example, human activities like tourists, commuting, and health care establish different kinds and levels of connections across the geographic units and thus form overlapping communities that may have shared local components [24]. In addition, from a community service perspective, people often reside in the service area of more than one public facility. For example, residents living in the middle of two clinics maybe accessible to both clinics. In this case, residents belong to the overlapping area of the two clinics' service coverages. If a disjoint community is carried out, residents can only be allocated to one of the two clinics. Therefore, overlapping community detection can effectively alleviate resource constraints and improve social services' efficiency. However, to the best of our knowledge, no overlapping community detection research has been conducted yet in the geography domain.

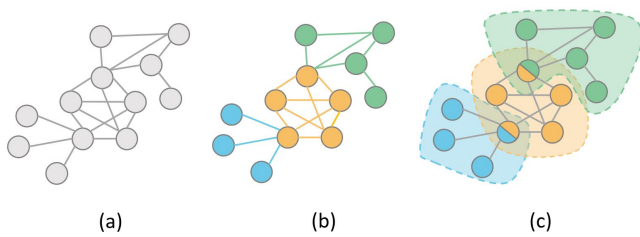


Figure 1: Spatial networks and communities: (a) Spatial network of geographical units; (b) Disjoint communities; (c) Overlapping communities

Furthermore, based on Granovetter’s theory, connections within networks can be complex, classified as both strong and weak ties [7]. Structurally embedded (tightly connected) edges are often socially strong, while long-range edges spanning different network parts are usually socially weak. Most existing community discovery algorithms assume that a set of nodes can only be treated as a community when they are more strongly connected compared to the expectation [17]. This basic assumption overlooks the fact that nodes with weak connections can structure communities that provide more information gains. For instance, it is acquaintances, not friends, who tell us more beyond our field [16]. Long-range and weak connections form critical community structures that can not be neglected for societal goods [23].

In sum, current community detection methods have not yet uncovered the overlapping nature of geospatial communities and has ignored long-range or weak ties in spatial networks. In this study, we provide a geospatial overlapping community detection (GOCD) framework based on the graph generation model and graph deep learning. First, we construct a spatial network based on the O-D information in human movements. Second, we proposed the Geospatial Graph Affiliation Generation model (GAGM), which generates the spatial graphs using the community affiliation probabilities and also provides the optimization objective for overlapping community detection. Third, we introduce the approach of graph convolutional networks to approximate the optimized community affiliation probabilities in the GAGM and thus uncover the geospatial overlapping community structures. In the remainder of the paper, we provide the literature review of the current community detection studies in Section 2. The proposed GOCD framework is described in Section 3. A case study and results are shown in Section 4. Finally, the study is concluded in Section 5.

2 RELATED WORK

Community detection. Traditional community detection methods mainly explore community from network structures [19], discovering non-overlapping communities, such as graph partition [4], statistical inference [9], and spectral clustering [2]. For example, with the goal of maximizing modularity, Louvain algorithm was invented for fast community discovery and has been widely used [3].

In order to discover overlapping communities, some methods based on the graph generation model (GGM) were proposed [20, 21]. The basic assumption of these methods is that two nodes will have a

stronger connection due to the more shared community affiliations. For example, people with more common hobbies are more likely to become friends. The first graph generation-based overlapping community detection model is the Community-Affiliation Graph (CAG) model [20]. It assumes that all nodes within a community are connected to each other with a fixed probability. However, the assumption is too strong in that it ignores the heterogeneity of nodes inside a community. To address this problem, the BIGCLAM (Cluster Affiliation Model for Big Networks) was proposed [21], which assumes that nodes are attached to the communities with different membership strengths. The membership strength determines the probability of having connections between any two nodes.

However, most traditional algorithms for overlapping community detection can only be applied to small networks. It is often difficult to achieve the desired results for complex, large-scale networks in the real world [19]. In recent years, applications of deep learning based on graph structures have achieved promising performances [18]. Many graph-based deep learning algorithms, especially the graph convolutional networks (GCN) have been applied to node prediction and link prediction tasks in various scenarios [12, 25]. Still, limited research has been done on overlapping community discovery in large networks. The Neural Overlapping Community Detection (NOCD) was proposed by combining the BIGCLAM and GCN [18], which was a pioneer attempt to achieve high detection accuracy for overlapping community structures and was applied to only non-spatial datasets.

Geospatial community detection. Geographic networks are more complex and have more nonlinear properties than most non-spatial networks, which pose a great challenge to overlapping community detection in geographic space. Compared with a large number of community detection studies in non-spatial scenarios (e.g., social networks), there are fewer studies on community structures in the geospatial context. Guo [8] constructs a geographic network based on trajectory data and combines multiple community structure measure metrics to perform community detection in human movements. Hong and Yao [10] implemented a hierarchical community detection for road networks based on the Infromap algorithm. Random walks and the Leiden technique were combined to detect the dynamical spatial community [11]. To the best of our knowledge, no studies of geographically overlapping community detection have been conducted.

3 METHODOLOGY

This study proposes the Geospatial Overlapping Community Detection (GOCD) framework. The basic idea of GOCD is to formalize the overlapping community detection tasks by discovering the most reasonable community affiliation matrix to reproduce the observed spatial networks. Our framework has two basic assumptions:

- Two geographical units are connected only when they belong to at least one geographical community.
- Within a geographic community, each one of the geographic units has a affiliation strength, which determines the probability of any two geographic units to be connected.

Specifically, GOCD consists of three steps: First, geospatial knowledge is introduced to construct spatial networks based on associations such as human movements. Second, we construct the Geospatial Graph Affiliation Generation model (GAGM), which can generate the whole network by the community affiliation information of each node. Since the community detection task is to get affiliation information from known networks, the GAGM model can be understood as the inverse task of community detection. Based on this, community detection can be understood as finding the affiliation matrix which has the highest possibility generates a known network. Hence, the GAGM provides an optimized objective function for our GOCD framework. Third, the GCNs is used to solve the community detection problem by discovering the optimized affiliation matrix. The computed community affiliation matrix provides the overlapped community memberships for each geographic unit.

Figure 2 illustrates the general process of detecting overlapping community structures from a geospatial network. We will explain each step in the following subsections. It should be mentioned that our model is inspired by the NOCD [18], which has been used for overlapping community detection in non-spatial networks. Since it is not designed for community detection with geographic contexts, spatial constraints and spatial associations are not considered in the NOCD.

3.1 Constructing spatial networks

The geographical units are interconnected to form a weighted graph $G = (V, E)$, where $V = \{1, \dots, N\}$ contains N geographic units, and $E = \{(u, v) \in V \times V : A_{uv}\}$ includes the connection of any two geographic units, where A_{uv} represents the weight of the edges between node u and node v . The connections can be distance, topological adjacency, human movements and many other types of geographic connections [25]. The weight of the edges between all nodes forms the adjacency matrix A . In addition, each node may have a vector of attributes in dimension D . The attribute vectors of all nodes form the attribute matrix $X \in \mathbb{R}^{N \times D}$, as shown in Figure 2(a)

3.2 Generating networks with community affiliations

Based on the ideas of CAG model [20] and BIGCLAM [21], we propose the Geospatial Graph Affiliation Generation model (GAGM). It can generate the geographic network by the community affiliation of each node.

Assume that the set of communities in a geographic network is C . There are two nodes m and k , and the membership affiliation strength vector of them are F_m and F_k , respectively. F_m is consist of the membership affiliation strength of node m to every community.

The GAGM create an edge (m, k) between nodes m and k with a probability $P(m, k)$:

$$P(m, k) = 1 - \exp\left(-F_m \cdot F_k^T\right) \quad (1)$$

The derivation process of Equation (1) is as follows. Suppose that there is a community c ($c \in C$) and the membership strength of nodes m and k to c is F_{mc} ($F_{mc} \in F_m$) and F_{kc} ($F_{kc} \in F_k$), respectively. The interaction strength between nodes m and k in

community c is $X_{mk}^{(c)}$, which we assume obey the Poisson distribution [18]:

$$X_{mk}^{(c)} \sim \text{Pois}(F_{mc} \cdot F_{kc}) \quad (2)$$

Then the total interaction strength X_{mk} is the sum of $X_{mk}^{(c)}$:

$$X_{mk} = \sum_c X_{mk}^{(c)} \quad (3)$$

Then it obeys the Poisson distribution:

$$X_{mk} \sim \text{Pois}\left(\sum_c F_{mc} \cdot F_{kc}\right) = \text{Pois}(F_m \cdot F_k^T) \quad (4)$$

Since the edge possibility $P(m, k)$ is same as $P(X_{mk} > 0)$:

$$P(m, k) = P(X_{mk} > 0) = 1 - P(X_{mk} = 0) = 1 - \exp\left(-F_m \cdot F_k^T\right) \quad (5)$$

Based on the above derivations, we know that using GAGM, if the membership strength of each geographic unit to each community is given, we can calculate the connection probability between any two nodes, and thus reproduce the overall geographic network (graph).

As the outcomes, we define the community affiliation matrix F , where F_{ij} denotes the community strength of the i -th node to the j -th community. Then the community detection task can be understood as the inverse of the following process: finding the optimal F that can generate the geographic network (G) constructed in the previous step with the maximum probability. That is, finding the F which can maximize the $P(G|F)$:

$$\begin{aligned} P(G|F) &= \prod_{(m,k) \in E} P(m, k) \prod_{(m,k) \notin E} (1 - P(m, k)) \\ &= \prod_{(m,k) \in E} \left(1 - \exp\left(-F_m^T F_k\right)\right) \prod_{(m,k) \notin E} \exp\left(-F_m^T F_k\right) \end{aligned} \quad (6)$$

3.3 Optimizing overlapped communities with Graph Convolutional Networks

Being different from traditional methods, we introduces GCN to solve the F matrix [18] in a graph-based deep learning manner:

- Define F as the output of GCN:

$$F := \text{GCN}_\theta(A, X) \quad (7)$$

- The objective of GCN is set as maximizing $P(G|F)$

Since the likelihood function involves a product of many small probabilities, we use the log likelihood as the GCN loss function:

$$\mathcal{L}(F) = \sum_{(m,k) \in E} \log\left(1 - \exp\left(-F_m^T F_k\right)\right) - \sum_{(m,k) \notin E} F_m^T F_k \quad (8)$$

In addition, the geospatial graphs are often extremely sparse, which means many pairs of nodes don't have human movements. In this case, the second term in equation (8) has a much larger contribution. We deal with this problem by balancing the two terms:

$$\mathcal{L}(F) = \frac{1}{|E|} \sum_{(m,k) \in E} \log\left(1 - \exp\left(-F_m^T F_k\right)\right) - \frac{1}{n^2 - |E|} \sum_{(m,k) \notin E} F_m^T F_k \quad (9)$$

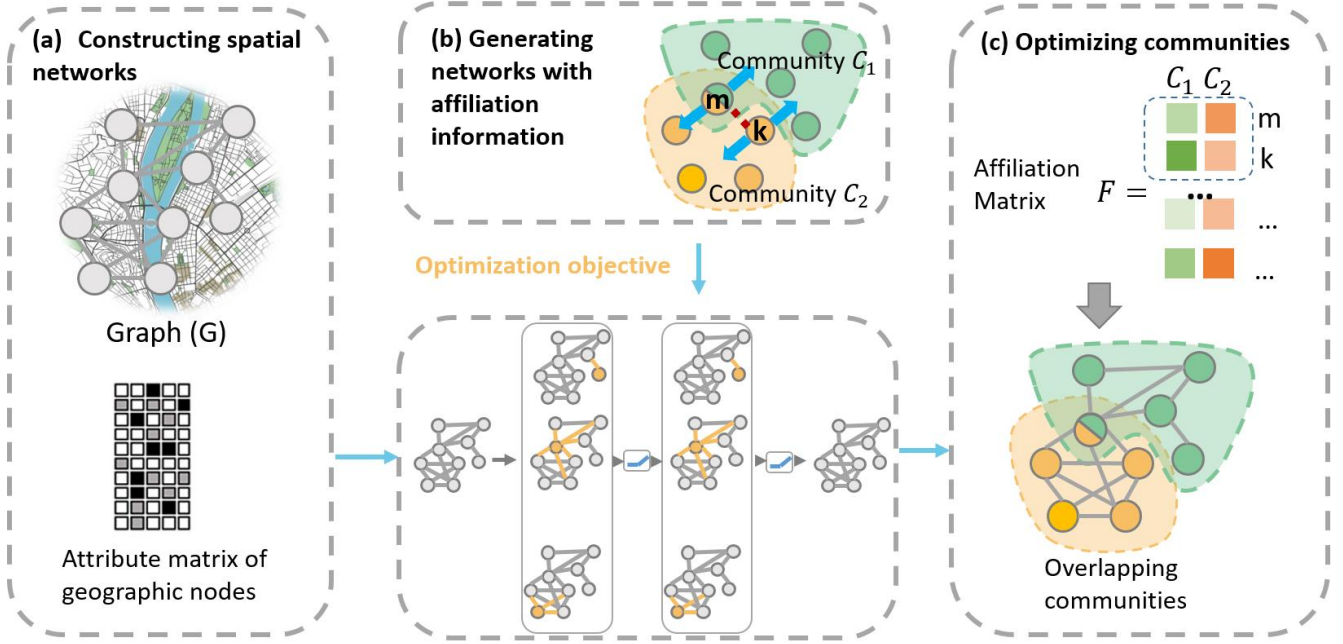


Figure 2: Illustration of the Geospatial Overlapping Community Detection framework

Thus, the optimization objective of the GCN is:

$$\theta^* = \arg \max_{\theta} \mathcal{L}(F) = \arg \max_{\theta} \mathcal{L}(GCN_{\theta}(A, X)) \quad (10)$$

After obtaining the optimized affiliation matrix F , we assign node m to community c if its affiliation strength F_{mc} is higher than the threshold β . The β is defined as the membership indicator which controls the connection strength of nodes inside the community. If β is 0, every pair of nodes with minimal weak connections will be assigned as the same community.

3.4 Model evaluation

Four metrics are used to evaluate the geospatial community detection result based on GOCD:

- Coverage (COV) describe what percentage of the edges is explained by at least one community. (i.e. if (u, v) is an edge, both nodes share at least one community). Higher coverage is better:

$$\text{Coverage}(C_1, \dots, C_K) = \frac{1}{|E|} \sum_{u, v \in E} \mathbb{1} \left[z_u^T z_v > 0 \right] \quad (11)$$

- Conductance (CON) is average conductance of the detected communities (weighted by community size). Lower is better.

$$\text{outside}(C) = \sum_{u \in C, v \notin C} A_{uv}$$

$$\text{inside}(C) = \sum_{u \in C, v \in C, v \neq u} A_{uv}$$

$$\text{Conductance}(C) = \frac{\text{outside}(C)}{\text{inside}(C) + \text{outside}(C)}$$

$$\text{AvgConductance}(C_1, \dots, C_K) = \frac{1}{\sum_i |C_i|} \sum_i \text{Conductance}(C_i) \cdot |C_i| \quad (12)$$

- Density (DEN) represents average density of the detected communities (weighted by community size). Higher is better.

$$\rho(C) = \frac{\# \text{ existing edges in } C}{\# \text{ of possible edges in } C}$$

$$\text{AvgDensity}(C_1, \dots, C_K) = \frac{1}{\sum_i |C_i|} \sum_i \rho(C_i) \cdot |C_i| \quad (13)$$

- Clustering coefficient (CC) describes average clustering coefficient of the detected communities (weighted by community size). Higher is better.

$$\text{AvgClustCoef}(C_1, \dots, C_K) = \frac{1}{\sum_i |C_i|} \sum_i \text{ClustCoef}(C_i) \cdot |C_i| \quad (14)$$

In addition, since geographic community detection is an unsupervised task with no community affiliation labels, we applied our model to a non-geospatial dataset for model validation, details can be found in Section 4.2.

4 CASE STUDY

4.1 Data and study area

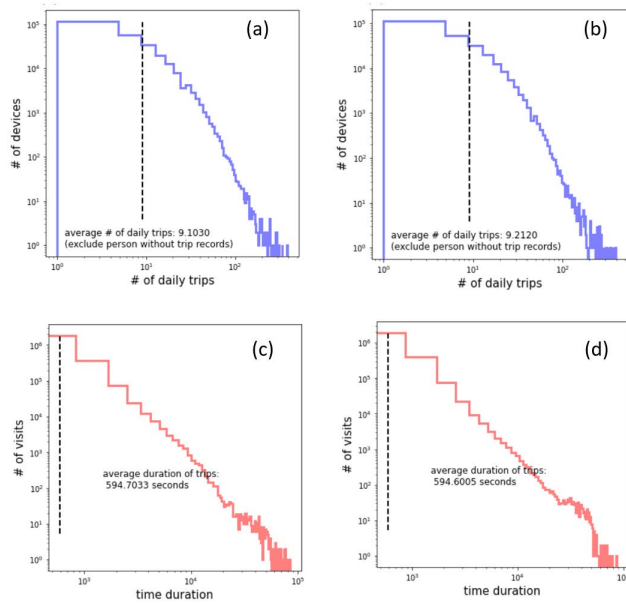


Figure 3: Statistic distribution of OD trips on a Monday (a,c) and a Sunday (b,d)

We conducted the overlapping community detection in the Twin Cities metro area (TCMA), Minnesota, U.S., to verify the effectiveness of our proposed framework. Minnesota is the largest state in the Midwest U.S. Its most important urban area is the Twin Cities region consisting of Minneapolis and St. Paul, which, along with their surrounding urban areas, account for more than half of Minnesota’s total residents. The Twin Cities contains seven counties, 186 CTUs, and 2085 census block groups (CBGs), according to the 2020 American Community Survey (ACS).

The trajectory data at the device level were collected from PlaceIQ¹, a location data and technology company for place intelligence. The raw data records the device ID, time, latitude and longitude information of the starting and ending points for each individual trip. We used two-day data for this study, i.e., 2021.03.01 (Sunday) and 2021.03.02 (Monday). The data for Sunday had a total of 2445310 trip records from 268627 devices, and the data for Monday had 2370828 trip records from 257363 devices.

We did a statistical analysis of the flow data, and the results are shown in Figure 3. On Monday, the average number of trips per device was 9.1030, and the average duration of each trip was 594.7033 s. On Sunday, the average number of trips per device was 9.2120, and the average duration of each trip was 594.6005 s. The average number of trips per device was 9.2120, and the average duration of each trip was 594.6005 s. On Monday, the average number of trips per device was 9.1030 and the average duration of each trip was 594.7033 s.

¹<https://www.placeiq.com/>

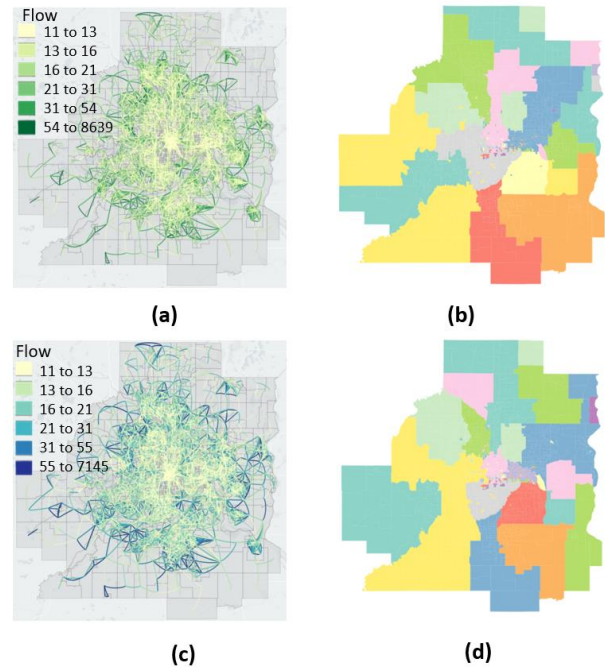


Figure 4: The CBGs-level human flows in Twin Cities Metro Area and the communities detected by Louvain method (Different colors represent different communities): (a) human flows on Monday, (b) communities on Monday, (c) human flows on Sunday, (d) communities on Sunday

We merge the human movement data into 2085 CBGs to get a 2085×2085 O-D matrix. Each OD_{ij} represents the number of flows from CBG i to j . The flow maps for the two days are visualized in Figure 4. In addition, we performed non-overlapping community detection using the Louvain algorithm (Figure 4 (b,d)). The results show that communities detected by the Louvain algorithm tend to exhibit spatial aggregation.

4.2 Spatial network construction and model settings

We define the edges between two geographic units in the graph as human flows to detect geospatial communities from such human flows. Therefore, a graph can be described as an adjacency matrix A , and A_{ij} represents the human flows between locations i and j . If we distinguish the departure and arrival of human flows, A_{ij} is the flow from i to j . In this context, the geospatial network is a directed weighted graph. If no distinction is made between departures and arrivals, then $A_{ij}=A_{ji}$ and the edge weight of the two geographic units is the sum of the flows between them. In this case, the geospatial network is an undirected weighted graph.

Since the relatively low intensity and long-distance flows may represent the trivial behaviours and cannot reveal the major patterns, we removed human flows less than 10, and flows with distance longer than 5 km. These two parameters were selected by checking the statistic distribution of the human flows and can be

further explored. After constructing the spatial networks, a classic two-layer GCN was adopted to optimize the GAGM [12, 18]. The process is defined as:

$$F := \text{GCN}_{\theta}(A, X) = \text{ReLU}\left(\hat{A} \text{ReLU}\left(\hat{A} X W^{(1)}\right) W^{(2)}\right), \quad (15)$$

where $\hat{A} = \tilde{D}^{-1/2} \tilde{A} \tilde{D}^{-1/2}$ is the normalized adjacency matrix introduced in [12].

Since there is no ground truth community affiliation label in our geospatial network, we optimized the architecture and membership indicators β using the benchmark dataset, inspired by the previous study. The benchmark dataset contains the co-authorship network developed from the Microsoft Academic Graph [18]. In this network, authors are graph nodes, and several research fields represent communities. The optimized β is found to be 0.5. In addition, we chose the community number $K=10$ via a sensitivity analysis. We set a series of K ranges from 1 to 100 and found ten to be the most reasonable. All the detected ten communities contain a reasonable number of geographical units (i.e. CBGs).

We verified the model performance in this non-spatial benchmark dataset, the results shown in Table 1 indicate that the trained model achieved high accuracy in detecting the overlapping communities.

Table 1: Community detection performance on the benchmark dataset

	COV	CON	DEN	CC
Ground truth	0.9588	0.3392	3.853E-03	-1.29E-04
Predicted	0.9240	0.2309	4.627E-03	-2.391E-05

4.3 Results

4.3.1 Spatial pattern of overlapping communities. The community detection was conducted using GAGM. Four indicators were used to describe the overlapping community properties, which are shown in Table 2. The results show overlapping communities have similar coverage and density on weekdays and weekends. Overlapping communities have 4.52% higher conductance on weekdays than at weekends. In addition, overlapping communities have a positive cluster coefficient on weekdays and a negative cluster coefficient on weekends.

Table 2: Community detection performance on weekday and weekend mobility networks

Time	COV	CON	DEN	CC
Weekday	0.4208	0.6246	1.077E-02	3.13E+01
Weekend	0.4187	0.5976	1.334E-02	-2.812E+01

The spatial distributions of the detected overlapping communities on a Monday (Figure 5) and a Sunday (Figure 6) were detected by the GAGM. The center maps in these figures represent the number of communities to which each CBG belongs (overlapping intensity).

Table 3: Spatial autocorrelation analysis of overlapping intensity

Time	Moran's Index	z-score	p-value
Weekday	0.1855	14.8523	0.0000
Weekend	0.1442	11.5547	0.0000

The results show that the downtown area has a higher overlapping intensity on Monday than Sunday. In addition, residential areas and shopping malls also have more places with higher intensity of community overlapping. On Sunday, the suburbs including some lake areas and golf clubs, have a higher overlapping intensity. The Minneapolis–Saint Paul International (MSP) Airport had the highest overlapping intensity. This indicates that on weekends, people from a wider area travel through the airport.

Further, we conducted a spatial autocorrelation analysis of overlapping intensity based on the global Moran's index. The results showed that both weekday and weekend overlapping intensity have significant positive spatial autocorrelation at $p < 0.01$, and the z-score of Moran's index was higher than 2.58. Among them, the z-score of Moran's index for Monday (14.8523) is higher than that of the Sunday (11.5547), which indicates that on a weekday, the overlapping intensity of the community shows a stronger spatial dependence: a geographic unit with strong flows is more likely to amplify the interactions of surrounding geographic units. This may be due to the fact that there are more human flows on a weekday between working places and residential areas. Meanwhile, on weekends, people travel for more diverse purposes, so the interactions between geographic units can be more complex and fragmented.

4.3.2 Local variations of the overlapping. We calculated the differences between the overlapping intensity of CBGs on Monday and on Sunday in Figure 7. A blue CBG means it belongs to more communities on Monday, and a red CBG means it belongs to more communities on Sunday. We selected four case regions for the exploratory analysis: (a) represents Minneapolis College of Art and Design, which belongs to 3 more communities on Sundays than on Mondays. This indicates that the college interacts more closely and diversifies with other areas during the weekend. The opposite is true for its surrounding residential communities, where there is higher community overlap on Mondays. (b) represents the campus of the University of Minnesota. (c) is located in a typical residential area. It has a much higher intensity of community overlap on Mondays than on weekends. This may be due to the fact that on weekdays, residents are out for working places, leading to a diverse interaction in these areas. (d) is the MSP airport, which has a lower community intensity on Monday than on Sunday. This may be due to the fact that there are more trips related to the airport on weekends.

4.3.3 Overlapping communities of the MSP airport. We selected the MSP airport for further interpretability analysis. As shown in Figure 7, we obtained the shared times of overlapping memberships with the airport for all CBGs. On Monday, the airport is subordinate to 3 communities, and on Sunday, it is subordinate to 5 communities. The average shared membership times to the MSP airport for all

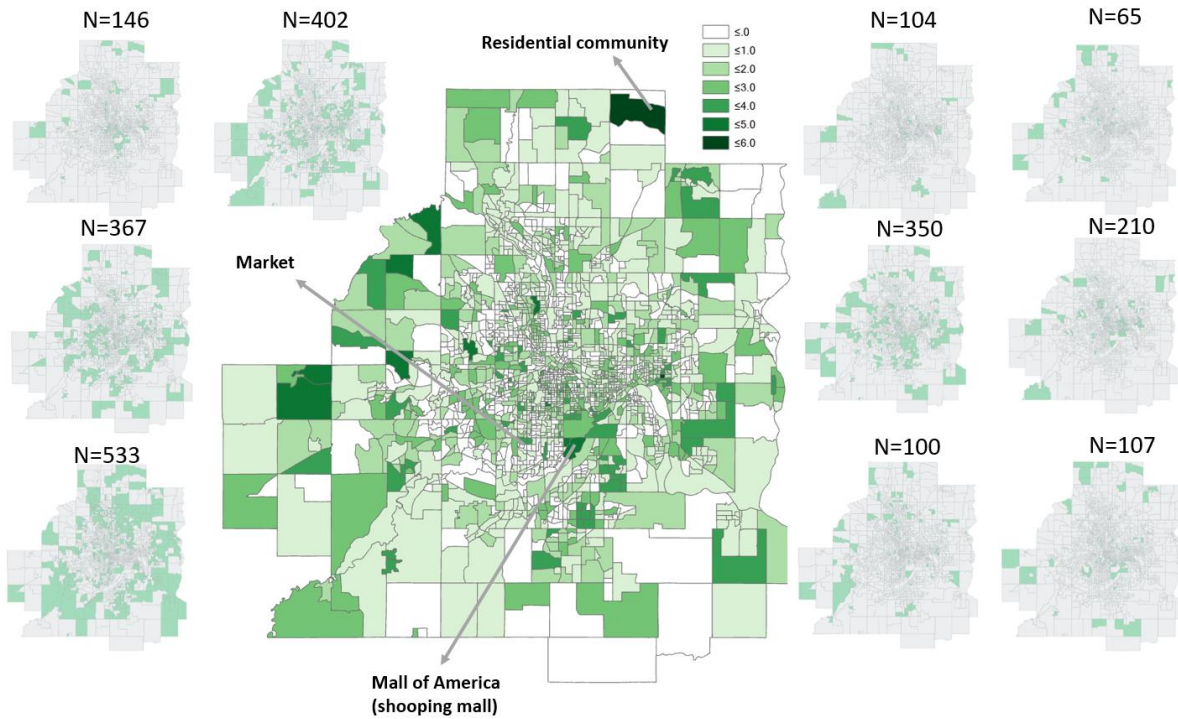


Figure 5: Overlapping community structure on Monday

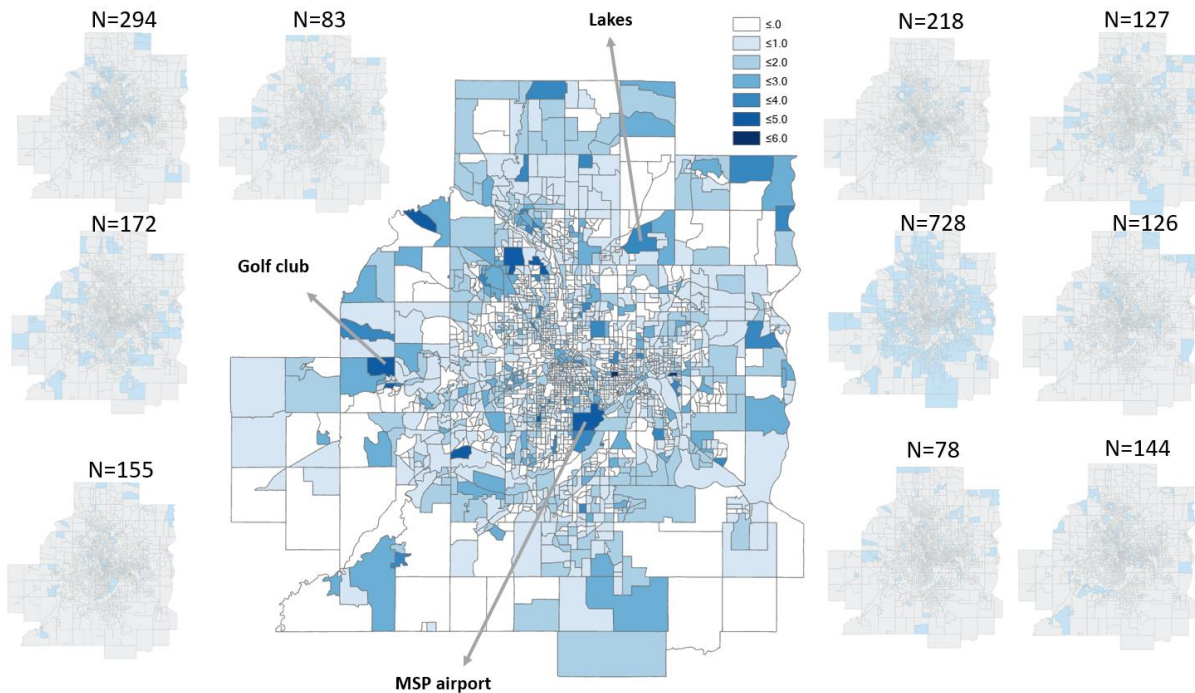


Figure 6: Overlapping community structure on Sunday

CBGs on Monday and Sunday is 0.6 and 0.65, respectively, indicating that on Sunday, the airport has a more complex overlapping community structure in terms of human movements.

We calculated the shared community frequency of all CBGs to the MSP airport. Results show that on Monday, 53.53 % of all

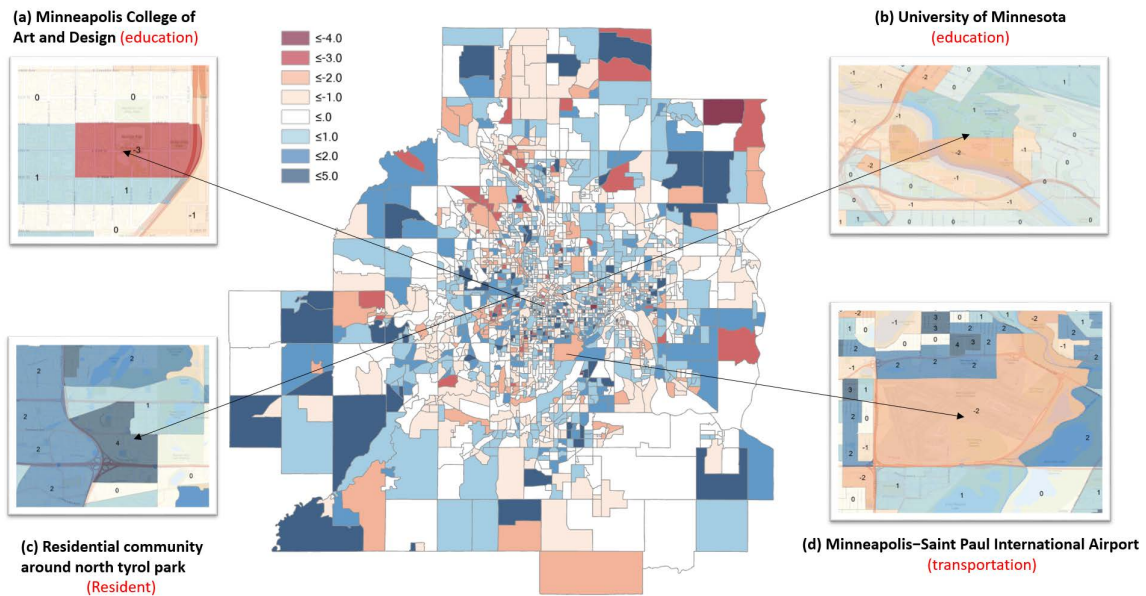


Figure 7: Overlapping intensity difference between Monday and Sunday. A blue CBG means it belongs to more communities on Monday, and a red CBG means it belongs to more communities on Sunday

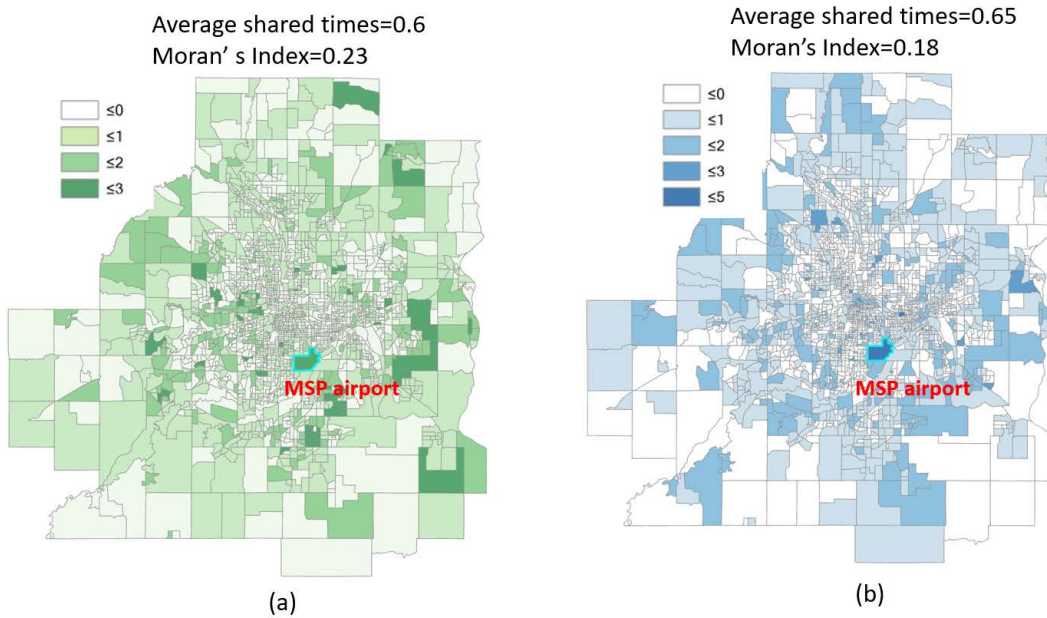


Figure 8: Shared times of community membership with MSP airport on (a) Monday and (b) Sunday

CBGs are not subordinate to any community that contains the MSP airport. While on Sunday, the proportion is only 48.63 %. In addition, on Monday, 918 CBGs have 1 or 2 times of community sharing with the MSP airport. On Sunday, the number becomes 1049. Notably, on Monday, 51 CBGs have three or more overlaps with the MSP airport, while there are only 19 on Sunday with three

or more overlaps. These findings indicate that although most areas have weaker connections to the airport on Monday compared to Sunday, there still exists a small number of areas with more shared communities to the airports on Monday, reflecting the importance of weak ties revealed by the geospatial overlapping community detection.

5 CONCLUSION

Communities in the real-world can be overlapped, yet there is no research focus on the overlapping community structures in geospatial networks. In this study, we combined a graph generation model with the graph convolution network for overlapping geospatial community detection from human movements. We collected trajectory data from the Twin Cities, MN for validation, and the results show significant spatial differences of overlapping community structures between a weekday and weekend. Also, overlapping communities have a positive cluster coefficient on the weekday and a negative cluster coefficient on the weekend. Further, we selected MSP airport for local analysis and found that compared to the weekday, the average overlapping intensity w.r.t. the airport is higher in the weekend, accompanied with a lower spatial autocorrelation in the weekend. Since the overlapping communities are not truly labelled, the community detection results can not be perfectly verified in this study. We will conduct more in-depth interpretability studies in the future to better validate the proposed GOCD framework. The GOCD framework can help to better describe the complex urban structure and explore the potential connections between local regions, guiding sustainable community planning and resource allocation in future cities.

6 ACKNOWLEDGMENTS

The research is supported by the Faculty Interactive Research Program from Center for Urban and Regional Affairs (no.1801-10964-21584-5672018), and the New Faculty Set-up Funding of College of Liberal Arts, University of Minnesota (no.1000-10964-20042-5672018).

REFERENCES

- [1] Emmanuel Abbe. 2017. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research* 18, 1 (2017), 6446–6531.
- [2] Arash A Amini, Aiyou Chen, Peter J Bickel, and Elizaveta Levina. 2013. Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics* 41, 4 (2013), 2097–2122.
- [3] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefevre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008, 10 (2008), P10008.
- [4] Santo Fortunato. 2010. Community detection in graphs. *Physics reports* 486, 3-5 (2010), 75–174.
- [5] Esther Galbrun, Aristides Gionis, and Nikolaj Tatti. 2014. Overlapping community detection in labeled graphs. *Data Mining and Knowledge Discovery* 28, 5 (2014), 1586–1610.
- [6] Michelle Girvan and Mark EJ Newman. 2002. Community structure in social and biological networks. *Proceedings of the national academy of sciences* 99, 12 (2002), 7821–7826.
- [7] Mark S Granovetter. 1973. The strength of weak ties. *American journal of sociology* 78, 6 (1973), 1360–1380.
- [8] Diansheng Guo, Hai Jin, Peng Gao, and Xi Zhu. 2018. Detecting spatial community structure in movements. *International Journal of Geographical Information Science* 32, 7 (2018), 1326–1347.
- [9] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. 1983. Stochastic blockmodels: First steps. *Social networks* 5, 2 (1983), 109–137.
- [10] Ye Hong and Yao Yao. 2019. Hierarchical community detection and functional area identification with OSM roads and complex graph theory. *International Journal of Geographical Information Science* 33, 8 (2019), 1569–1587.
- [11] Tao Jia, Chenxi Cai, Xin Li, Xi Luo, Yuanyu Zhang, and Xuesong Yu. 2022. Dynamical community detection and spatiotemporal analysis in multilayer spatial interaction networks using trajectory data. *International Journal of Geographical Information Science* (2022), 1–22.
- [12] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).
- [13] Da Kuang, Chris Ding, and Haesun Park. 2012. Symmetric nonnegative matrix factorization for graph clustering. In *Proceedings of the 2012 SIAM international conference on data mining*. SIAM, 106–117.
- [14] Pierre Latouche, Etienne Birmelé, and Christophe Ambroise. 2011. Overlapping stochastic block models with application to the french political blogosphere. *The Annals of Applied Statistics* 5, 1 (2011), 309–336.
- [15] Ye Li, Chaofeng Sha, Xin Huang, and Yanchun Zhang. 2018. Community detection in attributed graphs: An embedding approach. In *Thirty-second AAAI conference on artificial intelligence*.
- [16] Jianhua Ruan and Weixiong Zhang. 2006. Identification and evaluation of weak community structures in networks. In *AAAI*. 470–475.
- [17] Yiye Ruan, David Fuhry, and Srinivasan Parthasarathy. 2013. Efficient community detection in large networks using content and links. In *Proceedings of the 22nd international conference on World Wide Web*. 1089–1098.
- [18] Oleksandr Shchur and Stephan Günnemann. 2019. Overlapping Community Detection with Graph Neural Networks. *Deep Learning on Graphs Workshop, KDD* (2019).
- [19] Xing Su, Shan Xue, Fanzhen Liu, Jia Wu, Jian Yang, Chuan Zhou, Wenbin Hu, Cecile Paris, Surya Nepal, Di Jin, et al. 2022. A comprehensive survey on community detection with deep learning. *IEEE Transactions on Neural Networks and Learning Systems* (2022).
- [20] Jaewon Yang and Jure Leskovec. 2012. Community-affiliation graph model for overlapping network community detection. In *2012 IEEE 12th international conference on data mining*. IEEE, 1170–1175.
- [21] Jaewon Yang and Jure Leskovec. 2013. Overlapping community detection at scale: a nonnegative matrix factorization approach. In *Proceedings of the sixth ACM international conference on Web search and data mining*. 587–596.
- [22] Yao Yao, Jiale Wang, Ye Hong, Chen Qian, Qingfeng Guan, Xun Liang, Liangyang Dai, and Jinbao Zhang. 2021. Discovering the homogeneous geographic domain of human perceptions from street view images. *Landscape and Urban Planning* 212 (2021), 104125.
- [23] Fan Zhang, Jinyan Zu, Mingyuan Hu, Di Zhu, Yuhao Kang, Song Gao, Yi Zhang, and Zhou Huang. 2020. Uncovering inconspicuous places using social media check-ins and street view images. *Computers, Environment and Urban Systems* 81 (2020), 101478.
- [24] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. 2009. Mining interesting locations and travel sequences from GPS trajectories. In *Proceedings of the 18th international conference on World wide web*. 791–800.
- [25] Di Zhu, Fan Zhang, Shengyin Wang, Yaoli Wang, Ximeng Cheng, Zhou Huang, and Yu Liu. 2020. Understanding place characteristics in geographic contexts through graph convolutional neural networks. *Annals of the American Association of Geographers* 110, 2 (2020), 408–420.

A2. Spatial disparities in trade-offs: economic and environmental impacts of road infrastructure on continental level

Reference: Luo, P., Song, Y., & Wu, P. (2021). Spatial disparities in trade-offs: economic and environmental impacts of road infrastructure on continental level. *GIScience & Remote Sensing*, 58(5), 756-775.



Spatial disparities in trade-offs: economic and environmental impacts of road infrastructure on continental level

Peng Luo, Yongze Song & Peng Wu

To cite this article: Peng Luo, Yongze Song & Peng Wu (2021): Spatial disparities in trade-offs: economic and environmental impacts of road infrastructure on continental level, GIScience & Remote Sensing, DOI: [10.1080/15481603.2021.1947624](https://doi.org/10.1080/15481603.2021.1947624)

To link to this article: <https://doi.org/10.1080/15481603.2021.1947624>



Published online: 01 Jul 2021.



Submit your article to this journal [↗](#)





View related articles [↗](#)



View Crossmark data [↗](#)



Spatial disparities in trade-offs: economic and environmental impacts of road infrastructure on continental level

Peng Luo^a, Yongze Song ^b and Peng Wu ^b

^aChair of Cartography, Department of Aerospace and Geodesy, Technical University of Munich, Munich, Germany; ^bSchool of Design and the Built Environment, Curtin University, Perth, Australia

ABSTRACT

Remote sensing and geospatial techniques are being used to provide large-scale and regional solutions for achieving the sustainable development goals (SDGs) of the United Nations, including sustainable infrastructure development. Road transportation infrastructure has a significant contribution to the economy, but it also increases environmental pressure. However, little knowledge is available about spatial characteristics in the relationship between road impacts on the economy and impacts on the roadside environment. This research explores the spatial disparities in the relationship of road impacts on a continental level in Australia from 2011 to 2016. The performance of road transportation infrastructure is characterized from the perspectives of road density, connectivity, traffic volumes, and service to communities, other transportations (e.g. ports and airports), and industries, using remote sensing data and spatial heterogeneity models. Local economy and roadside environment are respectively presented using resident income and the change of roadside Enhanced Vegetation Index (EVI) and Aerosol Optical Depth (AOD) derived from the moderate resolution imaging spectroradiometer (MODIS) onboard the Terra satellite generated from Google Earth Engine. The road impacts of variables and their interaction on the economy and environment were calculated using an optimal parameters-based geographical detectors model (OPGD). Results reveal that the interaction of road density and traffic volumes can explain 47.4% of the resident income. In addition, results demonstrate the significant spatial disparities in the relationship between road impacts on the economy and impacts on the local environment. In major cities, such as Sydney and Melbourne, the pressure of roadside environment is increased with the economic growth, but the roadside environment has been improved in suburban and rural areas. Areas with the service to industries range from 64.4 km to 128 km have the most significant roadside EVI increase (2.5%). To the best of our knowledge, this is the first research to explore spatially differentiated trade-offs between the economic and roadside environmental impacts of roads using remotely sensed data, geospatial data, and spatial heterogeneity model at the continental level. Findings from this study provide an in-depth understanding of the interactions and trade-offs of road impacts on the local economy and the environment. Geospatial trade-offs and impact analysis methods in the study can be applied in wider fields to achieve global and regional SDGs.

ARTICLE HISTORY

Received 8 March 2021
Accepted 18 June 2021

KEYWORDS

Roadside environment; sustainable infrastructure development; MODIS EVI and AOD; Google Earth Engine; spatial heterogeneity; Australia

1. Introduction

Remote sensing and geospatial techniques provide effective data-driven solutions and opportunities for achieving the sustainable development goals (SDGs) of the United Nations. The remote sensing supported solutions for achieving SDSs generally include three categories. First, remote sensing can provide massive large-scale and timely Earth observation data for analyzing sustainability (Anderson et al. 2017; Cochran et al. 2020; Im 2020; Pathak et al. 2021). Remote sensing sensors can cover a large area with a rapid update frequency, making it possible to detect climate change, disasters, and health at a global scale

(J. Yang et al. 2013; Viana et al. 2017). Next, geospatial techniques are essential tools for investigating spatial and spatiotemporal patterns, exploring factors, and future scenario prediction. Long time-series remote sensing images are helpful for understanding mechanisms of human-environment interaction and effectively dealing with environmental challenges (Bishop-Taylor, Tulbure, and Broich 2018). Finally, data-driven solutions generated from remote sensing and geospatial techniques are solid and quantitative evidence for management and making practical decisions. The environmental variables, such as air quality variables (Alvarez-Mendoza, Teodoro, and Ramirez-

Cando 2019), obtained from remote sensing data are universally meaningful compared with in-situ and ground monitoring data due to the large spatial scale and long duration for effective decision-making (Boulila, Farah, and Hussain 2018).

Sustainable transportation infrastructure is one of the key sectors to improve economic development and social well-being among SDGs. A primary function of transportation infrastructure is to connect different regions, which is helpful for providing job opportunities and economic activity (Agbelie 2014). Well-performed transportation infrastructure also enables high accessibility to markets and raw materials and raises productivity due to the reduction in traffic congestion and travel time (Agbelie 2014; Umar et al. 2020). Given its importance on the social economy, hundreds of billions of dollars are spent on infrastructure investment and construction worldwide every year (Allen and Arkolakis 2020).

Road transportation infrastructure can provide great economic opportunities, but it may also increase local environmental pressure (Damania et al. 2018). The roadside environmental pressure caused by road transportation includes air pollution, soil erosion, vegetation and forest degradation, risks to species diversity, etc. Developed transportation infrastructure usually means high traffic volumes, leading to an increase in emissions, causing air pollution, and the urban heat island effect (Karagulian et al. 2015). Road transportation is closely associated with roadside soil pollution such as the increase of soil pH and heavy metals, and soil erosion (Jantunen et al. 2006; Ghosh, Raj, and Maiti 2020). Besides, the roadside environmental changes can decrease the natural growth of vegetation (Ghosh, Raj, and Maiti 2020), cause species diversity loss (Jantunen et al. 2006; Deljouei et al. 2018), lead to forest degradation (Mann, Agrawal, and Joshi 2019), and threaten the ecosystem. In order to reduce environmental pressure while safeguard economic growth, authorities usually face trade-offs of road impacts, which is the coordination between impacts on the economy and impacts on the environment. However, the methods and knowledge about identifying and understanding the trade-offs of road impacts are still limited.

The roadside environment can be characterized by combined ground monitoring data and remote-sensing data. Most of the earlier studies explored the relationship between environmental factors and

transportation investments using in-situ data, including soil samples for analyzing heavy metal pollution around roads (Ghosh, Raj, and Maiti 2020) and data from environmental monitor stations for investigating the environmental impacts of transportation infrastructure (Jantunen et al. 2006). However, monitoring stations dedicated to detecting the roadside environment are often relatively few and sparsely distributed spatially, making it difficult to conduct large-scale studies. Remote sensing has become a common data source to represent the roadside environments, and it is more effective in providing essential data for characterizing roadside environments than station-based monitoring data at a large spatial scale. Roadside environmental variables retrieved from remote sensing data include soil variables, climate variables, and vegetation variables. For instance, soil moisture (Al-Yaari et al. 2019), soil heavy metal content (Y. Ge, Thomasson, and Sui 2011) can be estimated using remote sensing technology, which has equivalent accuracy with situ observation (Ma et al. 2019). Aerosol Optical Depth (AOD), a key physical quantity characterizing the degree of atmospheric turbidity, is an important factor in determining aerosol climate effects and estimating environmental pollution levels (Martins et al. 2019). Vegetation indexes and LST data can be used to assess the road impacts on roadside vegetation and trees (Cârlan et al. 2020). Vegetation indexes calculated by different bands of remote sensing images are often used to reveal vegetation situations. Among them, the enhanced Vegetation Index (EVI) is a very commonly used vegetation index that can effectively reflect vegetation changes (Rashid Khan et al. 2018).

The interactive impacts of different variables of road infrastructure on the economy and environment are sophisticated, leading to the difficulty of quantifying trade-offs of road impacts (Allen and Arkolakis 2020). It is also a challenge to estimate regional disparities in the road impacts on the economy and the environment due to the ubiquitous spatial heterogeneity in both road performance variables and economic and environmental variables. Therefore, spatial heterogeneity methods are required to explore the road impacts on the economy and the environment and its trade-offs. The geographical detector model is an effective approach to investigate the spatial heterogeneity in the stratified structure of variables without the requirement of statistical

distributions of data (Y. Hu et al. 2011; Song, Wu et al. 2020a). It has been widely used in investigating environment change (Du et al. 2016; Shrestha and Luo 2017; Ding et al. 2019; Zhu, Meng, and Zhu 2020), urban expansion (Yang, Qian, and Long 2016), health risk assessment (J. F. Wang et al. 2010; Erjia et al. 2017; Liao et al. 2017), natural disaster risk assessment (Hu et al. 2011; Zhang, Nie et al. 2020). In recent studies, the geographical detectors model has been applied in transportation studies. For instance, it is used to identify influence factors of traffic accidents (Y. Zhang, Lu, and Qu 2020) and traffic jam (Daniel(Jian), Kaisheng, and Suwan 2018), explore the impact of transportation modes on epidemic (Cai et al. 2019) and impact of transportation on population distribution (L. Wang and Chen 2018), and analyze road deterioration (Song et al. 2018b, 2020b). An optimal parameters-based geographical detector (OPGD) was developed to explore the relationship between road deteriorations and potential explanatory variables, such as traffic volumes, climate, and soil attributes (Song et al. 2020b). Geographical detectors model is also widely used to study the association between road transportation and the roadside environment, such as traffic emissions (Daniel(Jian), Kaisheng, and Suwan 2018), heavy metal pollutions around the transportation hub (D. Li and Liao 2018), wildlife movements affected by roads (Shi et al. 2018), and build environment (S. Wang et al. 2018; Li, Lyu et al. 2020b).

In general, current research lacks a spatial analysis approach to explore the trade-offs between the impact of road infrastructure on the roadside environment and the economy. Thus, most of the road performance indicators and roadside environmental data used are statistical data and roadside environmental monitoring stations, making it difficult to provide sufficient spatial information. In this study, spatial disparities in the trade-offs between road impacts on the economy and the roadside environment were investigated with remote sensing and geospatial data using an optimal parameters-based geographical detector (OPGD) model. First, road performance was characterized from the perspectives of road density, connectivity, traffic volumes, and services to communities, other transportations (e.g. ports and airports), and industries. The services of road infrastructure were evaluated using a spatial accessibility analysis with OpenStreetMap (OSM) derived points of interest (POIs). Next, based on the Google

Earth Engine (GEE) platform, the change of roadside EVI and AOD were calculated and used to characterize the roadside environment. Resident income was used as a proxy variable of the local economics in this study. Third, an OPGD model was utilized to assess the spatial trade-offs of road impacts on the economy and roadside environment. In this step, optimal parameters of spatial discretization were derived for estimating the power of determinant (PD) and the power of interactive determinant (PID) of indicators of road infrastructure for the local economy and roadside environment. The nonlinearity and spatial disparities of impacts of individual road performance variables were compared to assess the spatial trade-offs. Finally, a sensitivity analysis was performed to evaluate the parameters of the roadside environment on the model and results. To the best of our knowledge, this is the first research to explore spatially differentiated trade-offs between the economic and roadside environmental impacts of roads using remotely sensed data, geospatial data, and spatial heterogeneity model at the continental level.

2. Study area and data

2.1. Study area

The transportation infrastructure systems are fundamental land assets in Australia. Australia has a well-developed road infrastructure consisting of 800,000 kilometers of road networks and transportation facilities, which is one of the most expanded road networks in the world. Transportation infrastructure makes a major contribution to the Australian economy in terms of employment, production, and exports, and contributed 7.4% to GDP in 2015–16 in Australia (ABS 2015). Australian authorities have developed national strategies on progressing toward transportation infrastructure-related SDGs, such as SDG 9 – building resilient infrastructure and SDG 11 – making cities and communities sustainable (Allen et al. 2019; Hall et al. 2020; Allen et al. 2020). In this study, the Local Government Areas (LGA) in Australia are used as the spatial unit of analysis.

2.2. Income data

Resident income data collected at the LGA level in 2011 and 2016 were used to demonstrate the local economy

in Australia (ABS 2020). The data is compiled from the Linked Employee Dataset (LEED), based on the Australian taxation system. The income values of the data represent personal income before any taxation, levies (e.g. Medicare levy), and losses, and inflation is not considered. In this study, the average income at each LGA between 2011 and 2016 was calculated to represent the local economy during this period.

2.3. Environment data

MODIS EVI data and MODIS AOD data were used to reveal roadside environmental changes related to road transportation infrastructure. The MOD13Q1 V6 product provides vegetation index values on a per-pixel basis with a resolution of 250 meters (Didan et al. 2015). The product contains two vegetation index bands, NDVI and EVI. The NDVI is the difference between near-infrared and red reflectance divided by the sum of them. The EVI is computed as follows:

$$EVI = G * \frac{\rho_{nir} - \rho_{red}}{\rho_{nir} + (C_1 * \rho_{red} - C_2 * \rho_{blue})} + L \quad (1)$$

where ρ_{nir} , ρ_{red} , and ρ_{blue} are near-infrared band, red band, and blue band of MODIS, respectively. G is the gain factor and equals 2.5 for MODIS EVI data. C_1 and C_2 are the coefficients of the aerosol resistance term, which are 6.0 and 7.5, respectively. L is the canopy background adjustment which equals 1.0 for MODIS EVI data. The value of EVI and NDVI are from -1 to 1 .

Compared to NDVI, EVI minimizes canopy background variation and maintains sensitivity in dense vegetation conditions (Matsushita et al. 2007). Also, EVI uses the blue band to eliminate residual atmospheric pollution caused by smog and sub-pixel thin-cloud cover. In addition, the MODIS Terra AOD (MCD19A2 V6) data with the resolution of 1 km is used as a proxy variable of roadside emissions (Martins et al. 2019). It contains AOD bands at 0.470 and 0.550 μm , where the 0.550 μm band of AOD is used in this study due to its wide applications in exploring environmental pollution issues (Allen et al. 2015).

2.4. Data of road performance

2.4.1. Traffic data

Traffic volume is a wide used indicator of the transportation infrastructure capacity. In this study, the

number of motor vehicles is used as proxy variables of nationwide traffic volume data across Australia. Australia bureau of statistics publishes the number of motor vehicles by LGA in 2006, 2011, and 2016. The mean volume at the LGA level in 2011 and 2016 was calculated as the average traffic volume during this period.

2.4.2. Facilities and networks

The data of facilities and road networks used in this study were derived from the OSM (<http://www.openstreetmap.org>). The OSM is the most widely recognized volunteered geographic information (VGI) data, and the database consists of vector data, such as facility location data, road networks, administrative boundaries, and land cover (Haklay and Weber 2008; Schultz et al. 2017). As is shown in table 1, in this study, the facilities data contain 28 types of POI at level B from nine types at level A: education facility, health facility, green spaces/sports area, public facility, residential area, airport, port, industry area, and commercial area (Table 1). Road networks from OSM in Australia contain multiple hierarchies of roads, and six hierarchies were selected to calculate road infrastructure performance, which are primary road, primary road link secondary, secondary link, trunk, and trunk link.

2.4.3. Population data

Spatial distributions of the population are characterized using the WorldPop population density data (<https://www.worldpop.org/>), which was generated to provide an open population dataset for sustainability development, disaster management, and health applications (Stevens et al. 2015b; Gaughan et al. 2013). The WorldPop data provide distributions

Table 1. Facility category and levels of POI.

Category of facility	POI at Level A	POI at Level B
Communities	Education facility	University, college, library, kindergarten, school
	Health facility	Hospital, pharmacy
	Green space/Sport area	Park, sport center, playground
	Public facility	Police, post office, fire station
	Resident area	Hotel, resident community
Other transportations	Airport	Airport
	Port	Port
Industries	Industry area	Waste, water plant, water tower, wind mill, factory
	Commercial area	Supermarket, bank, ATM, restaurant, cinema, theater, shop

of different kinds of population attributes, including density and age structure. The population density of WorldPop was mapped using a random forest model with a wide range of ancillary data and downscaled from the census data at an administrative level to the grid level (Stevens et al. 2015a; Gaughan et al. 2013). In this study, the WorldPop population density data at 1 km resolution in 2011 and 2016 were used for analysis.

3. Methods

Figure 1 shows the schematic overview of methods for assessing the spatial disparities in trade-offs between road impacts on the economy and the environment. In general, the methods consist of four steps: (i) Road density and road connectivity were calculated at an LGA level; (ii) spatial accessibility analysis was performed for each category of facilities, and the road services to communities, other transportations (e.g. ports and airports), and industries were estimated as the sum entropy weighted spatial accessibilities; (iii) roadside environment conditions were defined and quantified using MODIS EVI and AOD data. The roadside environment across the whole road network was computed on GEE. Fourth, PD and PID were estimated between road transportation infrastructure variables and economic or environmental variables. The trade-offs of road impacts on the economy and environment were analyzed through the comparison of their respective PD distributions; (iv) the sensitivity of the road buffer was explored by analyzing the road impacts on the environment at four different road buffers.

3.1. Calculation of road density and road connectivity

Road density and road connectivity are essential indicators to access road infrastructure and economic development. Road density in an LGA is a ratio between total length within the LGA and the area. Road connectivity is represented by the ratio of the number of interactions and the area. The original road network from OSM was segmented according to road attributes, such as road name, road hierarchy, and the max speed of the road. This segmentation may lead to massive junctions located in the middle of roads, and these junctions cannot reflect the transportation capacity. To address this issue, we merged the whole road network and then identified junctions at the intersections across the network. Thus, the density of identified junctions was used to present the road connectivity within LGAs.

3.2. Accessibility analysis

Road services to communities, other transportations, and industries were evaluated using a spatial accessibility analysis. Accessibility can be characterized by the average distance from the population to the facility. In this study, a network-based accessibility analysis was used to indicate services of the three types of facilities. The method to estimate the services of facilities includes the following steps.

First, population-weighted centroids (PWCs) were computed for LGAs. Due to spatial heterogeneity of population distribution, the PWCs can more accurately reveal the clustered location of the population

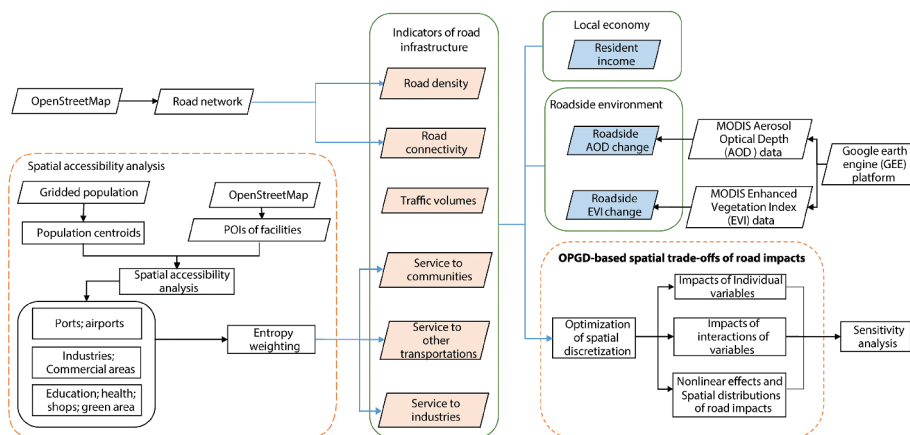


Figure 1. Schematic overview of assessing spatial disparities in trade-offs between road impacts on the economy and environment.

within a region than geometric centroids (Song et al. 2018a). The average population density between 2011 and 2016 was used to calculate PWCs using the PWC estimation method presented in Song et al. (2018a).

Second, spatial accessibility to the facility was computed using a network-based analysis method. POIs at level B were merged into level A, and all the PWCs and POIs were relocated in the road network according to the nearest Euclidean distance before the spatial accessibility. Accessibility of each category of the facility was represented by the average distance of the PWC to the nearest certain number of facilities. The numbers of target facilities were searched to determine the service capacity of facilities to local residents. If the number of targeted facilities is too small, the accessibility may be unstable and lack conviction, and if the number of target facilities is too large, the accessibility differences between different regions are blurred. A series of numbers of target destinations were set to select the suitable target destination count for different types of facilities. And the most reasonable destination counts were selected by visual checking. In this study, the number of target destinations for ports and airports and for other facilities was 2 and 10, respectively.

The final step was to present services of facilities using the sum of entropy-weighted spatial accessibility to facilities (Bao et al. 2020). Nine facility accessibilities were grouped into three categories, service to communities, service to other transportations, and service to industries using the entropy weight method. Entropy, an information indicator of variable, was used to evaluate the contribution of accessibilities for facilities in level A to the corresponding services to facilities. Thus, entropy was used to estimate the weights of accessibilities. If the information entropy of accessibility is high, a high weight should be given to the accessibility to a certain type of facility. The first step of the entropy weight method was to normalize the accessibilities. Then, the standardized value of the j th type of POIs at level B within the i th LGA is calculated as follows:

$$S_{ij} = \frac{P_{ij}}{\sum_{i=1}^n P_{ij}} \quad (2)$$

where n is the number of LGAs, P_{ij} is the normalized number j th POI at the i th LGA. Then, the information

entropy (E_j) and the information entropy weights (EW_j) of the j th type of POI are calculated as:

$$E_j = -\ln(n)^{-1} \sum_{i=1}^n W_{ij} \log_2 S_{ij} \quad (3)$$

$$EW_j = \frac{1 - E_j}{m - \sum_{j=1}^m E_j} \quad (4)$$

where m is the number of POI types in this category. Finally, the service of roads to a category of facilities at the i th LGA is computed as a sum entropy weighted accessibility to facilities in this category:

$$\eta_i = \sum_{j=1}^m EW_j * a_j \quad (5)$$

where a_j is the accessibility of j th category of facility.

3.3. Local economy and roadside environment

Average resident income was used to represent the local economy. The missing data at five LGAs (5/541) were filled using the inverse distance weighting (IDW) method (Donald 1968). To evaluate the roadside environment change, a 1 km buffer around the road network was generated to calculate the mean values of environmental variables within the buffer (Figure 2). Then, the roadside change of EVI and AOD of each LGA between 2016 and 2011 were calculated using the 1 km buffer from the GEE platform.

3.4. OPGD-based spatial trade-offs analysis

PD and PID are used to investigate the impacts of road performance on the economy and environment. PD and PID were widely used indicators to represent the impact of explanatory variables on response variables from the perspective of spatial heterogeneity (J. F. Wang et al. 2010). PD is used to explain the impact of individual variables, and PID is used to explain the interactive impact of variables. The fundamental assumption of the indicators is that: if an explanatory variable has a significant influence on a response variable, they are probably distributed in similar spatial patterns. In this study, a strategy of the optimization of spatial discretization was used to identify optimal discretization parameters of road performance

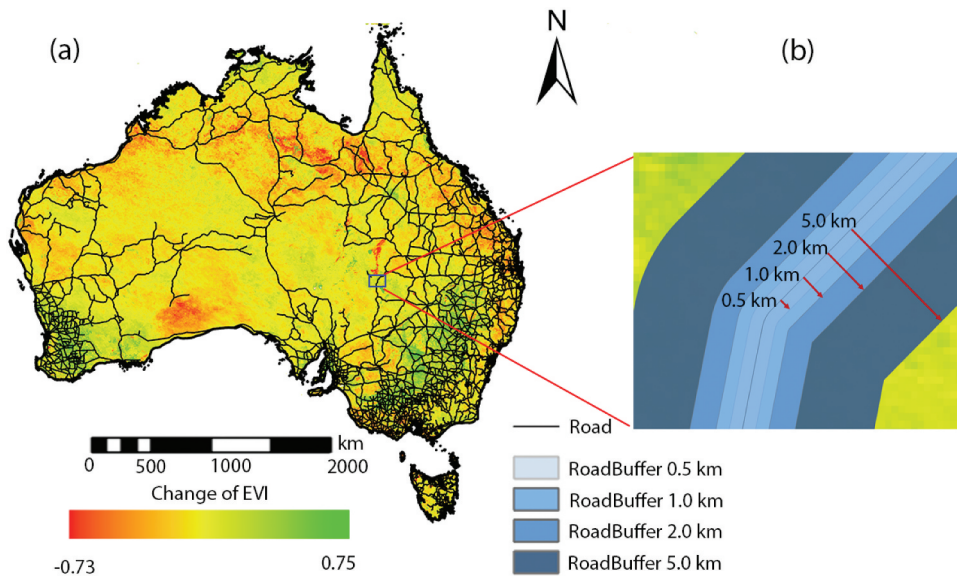


Figure 2. Road network and change of EVI derived from MODIS (a) and buffers for capturing roadside environment (e.g. EVI) (b).

indicators (Song et al. 2020b; Song and Peng 2021). Then, the PD and PID values were calculated to represent the impacts of six road performance indicators on resident income, change of roadside EVI, and change of roadside AOD, respectively. Finally, a sensitivity analysis was performed to validate the developed methods in the study. The optimization of spatial discretization was performed using the R package “ISDA” (Song and Peng 2021) and the PD and PID values were calculated using the R package “GD” (Song et al. 2020b).

3.4.1. Optimization of spatial discretization

The geographical detector model can only deal with discrete variables to calculate the PD and PID. Thus, all continuous variables need to be converted into discrete variables before inputting the model (Wang and Chengdong 2017). The spatial data discretization method aims to divide continuous geographical and geospatial data into several intervals depending on the data’s physical or statistical characteristics (Song et al. 2020b). In this study, the optimal spatial discretization method proposed by Song and Peng (2021) was used. Firstly, all variables are divided into 3–22 groups using a quantile break. Second, for each optional parameter combination of an explanatory variable, the Q value was calculated, and a variation curve of the 75th quantile Q values was smoothed using a locally estimated scatterplot smoothing (LOESS) model. Finally, when the increase rate of the

curve is lower than 5%, the point was selected as the optimal break number and used in further analysis.

Using the spatial discretization, the whole area was divided into several spatial overlay zones according to each explanatory variable. The average risk value at a zone of the explanatory variable was represented by the average value of response variables at this zone (Z. Wang et al. 2020). In this study, the spatial distribution of impacts of roads to the local economy and the roadside environment is assessed and visualized using the mean risk value.

3.4.2. Power of determinant

The PD is used to explore the explanatory power of road performance indicators on the economy and environment (Song and Peng 2021). The PD is measured by a Q value defined as:

$$Q = 1 - \frac{\sum_{z=1}^H N_z \sigma_z^2}{N \sigma^2} \quad (6)$$

where z is the number of spatial zones, N_z and N are the number of LGAs in zone z and the whole study area, respectively, and σ_z^2 and σ^2 are the variance of the response variable for the units in zone z and the whole study area, respectively. The Q value ranges from 0 to 1, and the Q value indicates that the explanatory variable explains $100 \times Q\%$ of the response variable. The Q value followed the noncentral F-test, which was used to determine the significance level (J. F. Wang et al. 2010)

3.4.3. Power of interactive determinant

The PID explains whether the explanatory powers of two factors are enhanced, weakened, or independent of each other (J. F. Wang et al. 2010; Hu et al. 2011; J. F. Wang, Zhang, and Fu 2016). First, the PD values of two explanatory variables X_A and X_B for the response variable were calculated as $Q(X_A)$ and $Q(X_B)$, respectively. Then, the PID of the interaction, a spatial overlay of factors X_A and X_B , was calculated as $Q(X_A \cap X_B)$. The comparison between PID and individual PD indicates if variables are spatially independent, enhanced, or weakened by each other. For instance, if $Q(X_A \cap X_B)$ is higher than the sum of $Q(X_A)$ and $Q(X_B)$, the interaction of X_A and X_B has an enhanced effect on the response variable. Conversely, the interaction of X_A and X_B has a weakened effect on the response variable. If $Q(X_A \cap X_B)$ equals to $Q(X_A) + Q(X_B)$, X_A and X_B are spatially independent when affecting the response variable.

3.4.4. Sensitivity analysis

The sensitivity analysis was conducted to explore the influence of parameters for defining roadside environment on the road impact assessment. In the study, the roadside environment was defined as EVI and AOD values at a 1-km buffer of roads. To evaluate the impacts of the distance of buffer on the road impact assessment, the impacts were calculated and compared for roadside EVI and AOD with four

different buffers around the road network in Australia (Figure 2b), including 0.5 km, 1 km, 2 km, and 5 km. The roadside change of EVI and AOD between 2011 and 2016 was evaluated within each road buffer based on the GEE platform. The PD values of road impacts on roadside environment variables at four buffers were calculated to analyze the sensitivity of the buffer distance on the impact evaluations. As a result, the variations of PD values under different distances of buffers can demonstrate the sensitivity of the methods.

4. Results

4.1. Spatial patterns

4.1.1. Local economy and roadside environment

Figure 3 shows spatial distributions and urban-rural comparisons of resident income and changes of roadside EVI and AOD. Resident income has a slight urban-rural disparity, where the urban resident income (\$59,299 per year) is 15.17% higher than the rural resident income (\$51,487 per year). Resident income in major cities is higher than in other regions.

From the perspective of the environment, the change of the roadside environment from 2011 to 2016 also contains regional disparities. Vegetation plays an important role in both the regional hydrological cycle, climate regulation, and ecological sustainability. EVI can characterize vegetation cover

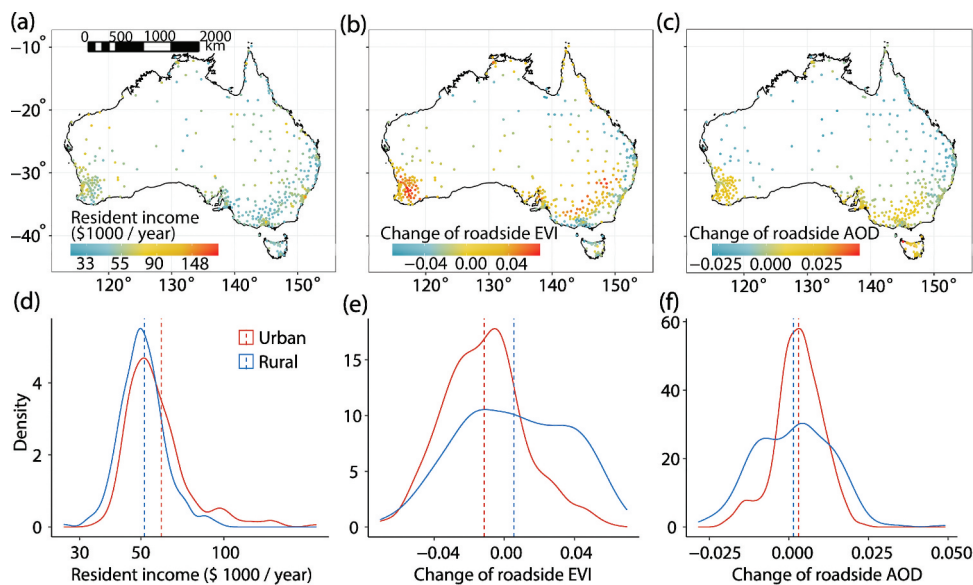


Figure 3. Spatial distributions (top) and density distributions (bottom) of local economy and environment changes: (a, d) resident income; (b, e) change of roadside EVI; and (c, f) change of roadside AOD. Roadside EVI and AOD are derived from MODIS.

changes very well (Duarte et al. 2018). The roadside EVI is increased in rural areas with 0.0058 EVI change, but it is decreased in urban areas with -0.0110 of EVI change. Roadside EVI decreases more in major cities than in suburban and rural areas. Suburban areas have the highest increase of EVI. The change of roadside EVI reflects the impact of human activities on the roadside environment, including infrastructure construction and road transport (Jantunen et al. 2006). The decrease of roadside EVI in urban areas reveals road transportation infrastructure has a negative impact on the environment (Ghosh, Raj, and Maiti 2020).

AOD is an indicator of atmospheric conditions in a region. The main factors contributing to the increase in AOD are polluting gases from industrial production, construction, transport, and other human activities. In Australia, AOD slightly increased in most areas, where the growth of AOD in urban areas is 2.1 times higher than that in rural areas, which was 0.0030 and 0.0014, respectively. The highest increase in AOD appears in the major cities and suburban areas. In inland areas and some parts of east areas, the AOD decreased, which indicates the improvement in air quality. In summary, from the spatial perspective, roadside environmental changes are closely linked with economic growth.

4.1.2. Road performance

Table 2 shows the weights of accessibilities to nine types of facilities in the three categories. The spatial

Table 2. The entropy weights of accessibility.

Accessibility to three categories of facilities	Accessibility different types of facilities in the category	Entropy weight
Communities	Residential area	0.227
	Public	0.197
	Health	0.180
	Green space and Sport area	0.224
	Education	0.172
Other transportations	Ports	0.518
	Airports	0.481
Industries	Industry	0.391
	Commercial	0.610

accessibility to each of the three categories of facilities, which is used to estimate the road services to facilities, is the sum of weighted accessibilities to different types of facilities in the category.

Figure 4 shows spatial distributions of road performance indicators. In general, road performance indicators perform better in major cities than that in suburban areas and rural areas. In major cities, services to communities and industries perform better than the service to other transportations, including ports and airports. The average distance of services to communities and industries ranges from 1.0 km to 7.4 km, while its ranges from 7.4 km to 54.6 km to other transportations.

The three direct road performance indicators, road density, road connectivity, and traffic volumes, are generally correlated with regional economic development. The differences between the three indicators in major cities, suburban areas, and rural areas are significant. As for the three indirect indicators, services to

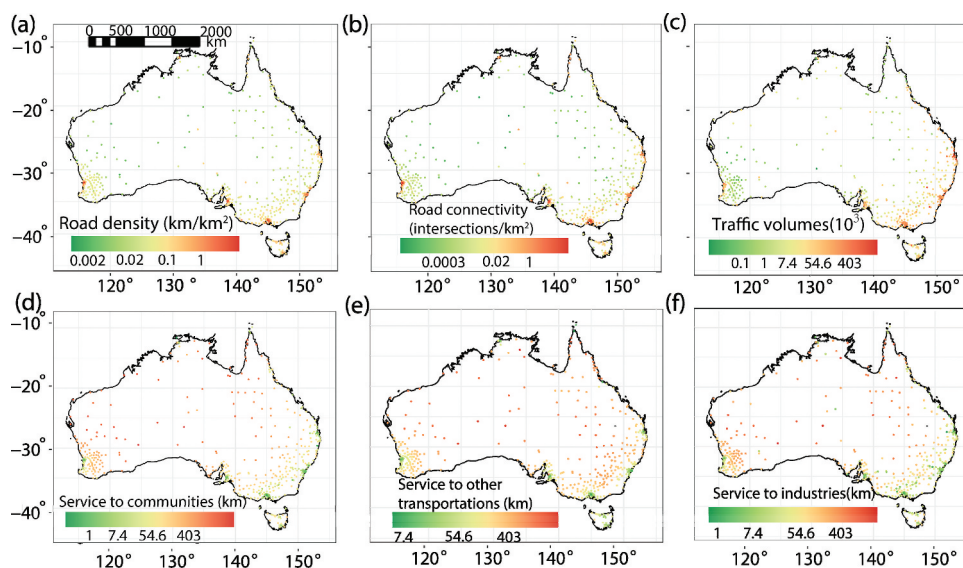


Figure 4. Spatial distributions of road performance indicators: (a) road density, (b) road connectivity, (c) traffic volumes, (d) service to communities, (e) service to other transportations, and (f) service to industries.

Table 3. Statistical summary of economic and roadside environmental variables and road performance indicators.

	Variable	Code	Mean	Median	SD	Min	Max	CV
Response variable	Resident income (\$1,000)	/	55.10	51.71	16.44	26.35	216.01	0.30
	Change of roadside EVI ^{**}	/	-0.0023	-0.0051	0.0290	-0.0704	0.0699	-12.41
	Change of roadside AOD ^{**}	/	0.0022	0.0027	0.0100	-0.0283	0.0490	4.54
Road performance indicators	Road density (km/km ²)	rd	0.44	0.08	0.84	0.00	5.76	1.90
	Road connectivity (interactions/km ²)	rc	1.07	0.02	2.77	0.00	22.01	2.58
	Traffic volumes (10 ³)	vlm	16.17	5.17	30.16	0.02	412.55	1.87
	Service to communities (km)	sc	91.97	43.13	136.54	0.85	1107.01	1.49
	Service to other transportations (km)	st	252.23	183.02	269.30	0.00	2096.52	1.07
	Service to industries (km)	si	55.39	22.78	91.62	0.00	870.30	1.65

^{**}Roadside EVI and AOD are derived from MODIS.

facilities, in addition to being influenced by the overall level of economic development, are also related to the industrial structure of the region and the type of facilities it leads to. For example, the services to other transportations in major cities and coastal cities ranges from 7.4 km to 54.6 km, while are above 54.6 km in suburban areas. However, in most of the suburban areas, services to the industry are below 7.4 km, with less difference with major cities.

Table 3 shows a statistical summary of response variables, including economy, roadside environment, and road performance indicators. The mean LGA-based income is 55,100, USD and the coefficient of variation (CV) of resident income is 0.30. The CV values of changes of roadside EVI and AOD are -12.87 and 4.54, respectively. The CV values indicate that the changes of EVI and AOD have much higher spatial disparities than resident income. The mean road density and connectivity are 0.44 km/km² and 1.07 interactions/km². The mean traffic volume of

LGAs is 16,170 per road. The mean distance between PWCs and facilities of communities, other transportations, and industries are 91.97 km, 252.23 km, and 55.39 km, respectively.

4.2. Road impacts on economy and environment

4.2.1. Optimal spatial discretization

Figure 5 shows the process of optimal spatial data discretization for the analysis of resident income, change of roadside EVI, and change of roadside AOD. With the break number increased from 3 to 16, the Q values of road performance indicators are generally increased (Figure 5 a–c), but the increase rate is gradually reduced (Figure 5 d–f). When the increase rate is lower than 0.05, the optimal break number is selected (Song and Peng 2021). The optimal numbers of spatial discretization are 7, 8, and 8 for spatial analysis of resident income, the change of EVI, and the change of AOD, respectively.

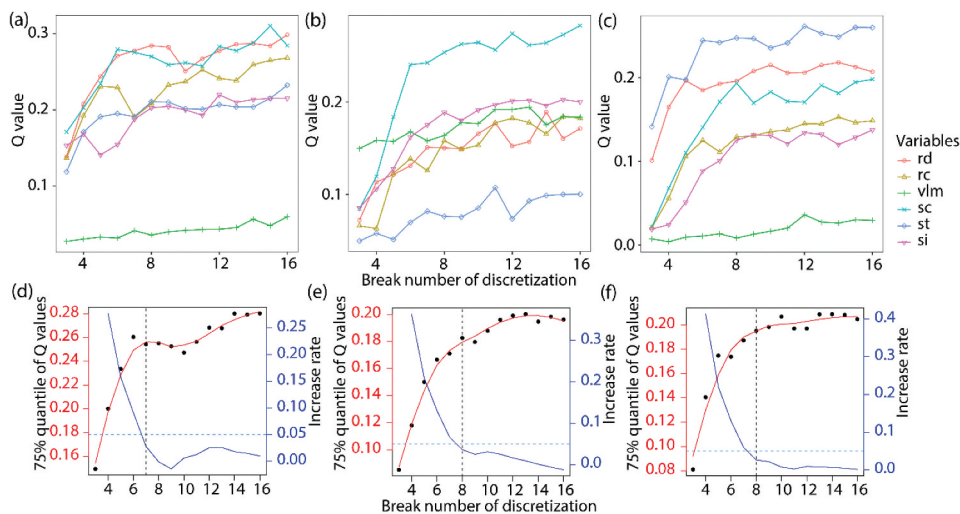


Figure 5. Processes and results of the optimization of spatial discretization for resident income (a, d), change of roadside EVI (b, e), and change of roadside AOD (c, f).

4.2.2. PD and PID of roads to economy and environment

Figure 6 shows the PD and PID of road transportation infrastructure on the local economy and roadside environment. Road density and service to communities have the largest impact on resident income, with Q values of 0.2775 and 0.2754, respectively. Highly dense roads can benefit the resident income. Road service to communities has a higher impact on resident income than services to other facilities. Community facilities include schools, hospitals, and public facilities, which are most relevant to human's daily activities. Results also show that the road service to public facilities brings more benefits to the resident income than the service to industries. Thus, the investment and construction of the living facilities are helpful for improving local resident income.

The interaction of traffic volumes with all the other variables has a nonlinear enhancing effect on resident income. Resident income is determined by multiple mixed factors, which are difficult to be measured (Glaeser, Kahn, and Rappaport 2008; Xue ting et al. 2018). This study shows that the interaction between traffic volumes and road density can explain 47.4% of resident income. The service to communities has the highest impact on the change of roadside EVI, with a Q value of 0.2539. And service to other transportations has the highest impact on the change of

roadside AOD, with a Q value of 0.2475. The interaction of service to other transportations and traffic volumes has the highest impact on the roadside environment, explaining 41.1% of the change of roadside EVI and 43.2% of the change of roadside AOD. Most of the imports and exports rely on port transportation in Australia. Maritime exports in 2016 were 1,394.5 million tonnes, comprising 909.5 million tons of crude oil and inedible materials (except fuels) and 440.2 million tons of mineral fuels, lubricants, and related materials in Australia (BITRE 2018). The change of the roadside environment is mainly because of the air pollution from freight transportation between ports and industrial regions, including mining, oil and gas products, grain, and other agricultural products.

Results show the distinctive impacts of traffic volumes on the local economy and roadside environment. Traffic volumes can be regarded as the proxy of economic vitality (Li, Gao et al. 2020a), while they can only explain 4.15% of resident income and 16.39% of roadside EVI. But our result shows traffic volumes and other road infrastructure performance have an extremely nonlinear enhanced impact on the local economy and roadside environment, which can explain nearly 50% of resident income, change of roadside EVI, and change of roadside AOD. This means only enough traffic volumes or economic vitality is insufficient to promote economic growth and

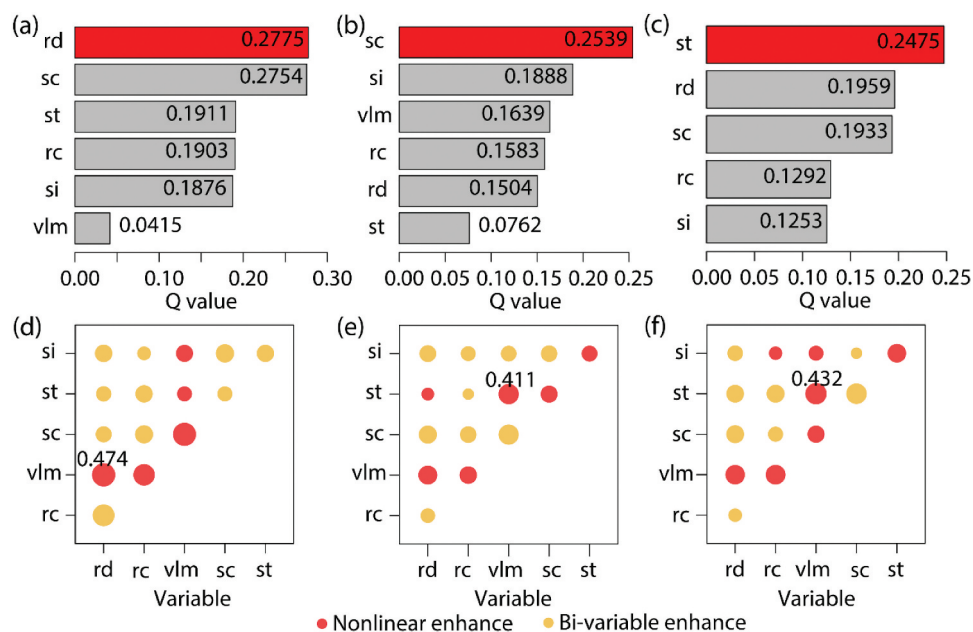


Figure 6. PD (top) and PID (bottom) of road performance indicators to resident income (a, d), change of roadside EVI (b, e), and change of roadside AOD (c, f).

environmental change, but when it is integrated with well-built road infrastructure, including roads and service facilities, the economy will develop rapidly, although environmental change will also intensify. On the other side, economic development also relies on adequate social-economic vitality, including a stable investment environment and reasonable economic policies. Only well-constructed infrastructure construction is not enough.

4.2.3. Spatial distributions and trade-offs of impacts

Road density and road connectivity are direct indicators to represent road performances. Figure 7 shows the impacts of road density and road connectivity on

the local economy and roadside environment, revealed by risk values from the geographical detector. The spatial disparities and nonlinearity of road impacts on the economy and environment are revealed. In major cities, road density and road connectivity have the highest impacts on resident income (refer to the red dots in the first map of Figure 7 a, b). Areas with the highest road density (range from 1.16 km/km² to 5.76 km/km²) and the highest road connectivity (range from 2.27 interactions/km² to 22 interactions/km²) have the highest resident income, which is 75,560 USD and 69,780, USD respectively. Resident income in suburban areas has the lowest dependence on road density and road connectivity.

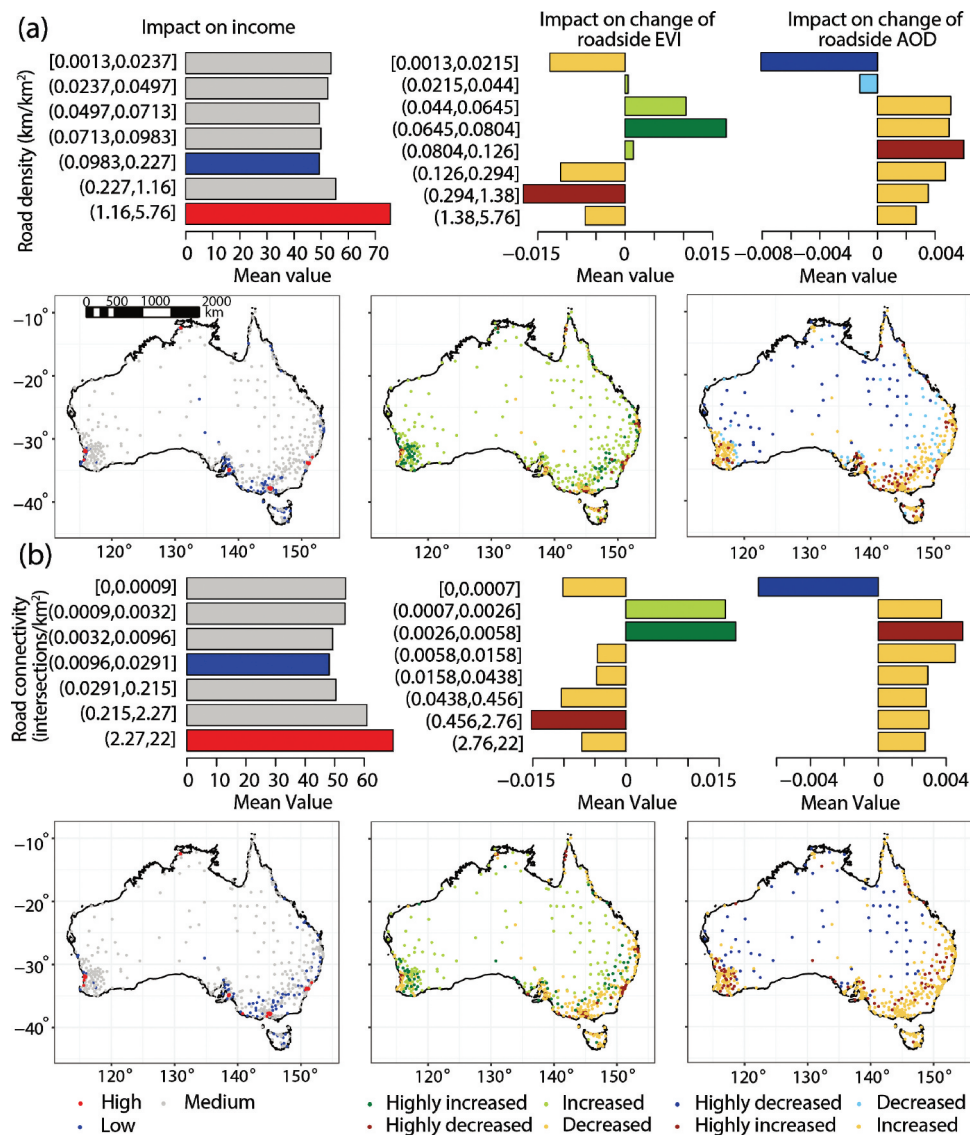


Figure 7. Trade-offs between road impacts on income and roadside environment: regional impacts of road density (a) and connectivity (b). Different colors of dots in maps represent the positive or negative impacts on the economy and environment. Roadside EVI and AOD are derived from MODIS.

As for environmental impacts, road density and road connectivity negatively impact roadside vegetation in major cities, with the decreased EVI (refer to the orange dots in the second map of Figure 7 a, b). The highest decrease EVI (−1.7%) appears in the areas with a road density range from 0.294 km/km² to 1.38 km/km². The second highest decrease of EVI is −1.5% in areas with road connectivity range from 0.456 interactions/km² to 2.76 interactions/km². In suburban and rural areas, road performance has a positive impact on roadside vegetation, especially in suburban areas. The highest increase EVI (1.8%) appears in the areas with road connectivity range from 0.0026 interactions/km² to 0.0058 interactions/km². EVI also increase a lot (1.7%) in areas with road density range from 0.0645 km/km² to 0.0804 km/km². EVI is a remote sensing indicator of vegetation coverage and vegetation condition. Results also indicate that the effects of protecting and recovering roadside vegetation are varied across different regions. Among different regions, actions in suburban areas have the highest positive effect on roadside vegetation growth and recovery. Therefore, the spatial disparities of road impacts on roadside vegetation should be considered in practical road asset management and decision-making.

Traffic volumes and service to facilities are indirect indicators for road performance. Figure 8 shows the impacts of four indicators on the local economy and

roadside environment. The findings from Figure 8 can be summarized into economic impact and environmental impact, as follows:

From the economic impact perspective, resident income in major cities has the highest dependence on transportation infrastructure performance (refer to the red dots in the first map of Figure 8 a-d). In suburban areas, services to facilities have relatively less importance to resident income (refer to the blue dots in the first map of Figure 8 b-d). Suburban areas with services to communities range from 33.7 km to 52.7 km have the lowest resident income (\$47,485). Therefore, the investment and construction of service facilities are usually helpful for the economic growth in suburban areas. In rural areas, traffic volumes have the lowest impact on resident income (refer to the blue dots in the first map of Figure 8 a). Thus, increasing traffic volumes in rural areas is a potential approach to stimulate the local economy since traffic volumes can partially indicate local socio-economic vitality.

From the environmental impact perspective, in major cities, roads have a negative impact on roadside vegetation, as demonstrated by the decreased EVI (refer to the orange dots in the second map of Figure 8 a-d). Areas with traffic volumes range from 37,400 to 413,000 have the highest reduction in roadside EVI (−1.7%). In rural areas, traffic volumes and services to facilities are beneficial to the roadside environment, increasing roadside EVI (refer to the

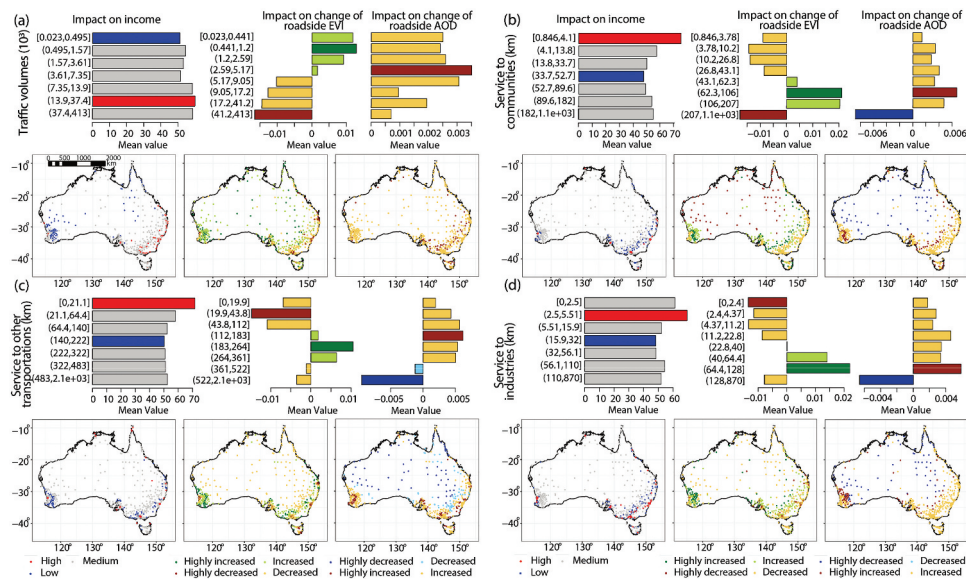


Figure 8. Regional impacts of traffic volumes and service to facilities. (a) Traffic volume; (b) service to communities; (c) service to other transportations; (d) service to industries. Different colors of dots in maps represent the positive or negative impacts on the economy and environment. Roadside EVI and AOD are derived from MODIS.

green dots in the second map of Figure 8 a). And the EVI in suburban areas increases more than in rural areas. Areas with service to industries range from 64.4 km to 128 km have the highest increase in roadside EVI (2.5%). Besides desert areas, all variables positively impact on the increase of roadside AOD, and traffic volumes have the highest impact. The increase of AOD in suburban areas is higher than that in major cities and rural areas.

To sum up, this study reveals the significant non-linearity, spatial disparities, and interactions of trade-offs between road impact on the economy and the local environment. In major cities, road-related income is much higher than that in other regions. On the contrary, the local economy in suburban areas has the lowest dependence on road performance.

Major cities with the highest road-related income have higher environmental pressure than other regions, which reveals interactions of trade-offs between road impact on the local economy and the roadside environment. In suburban and rural areas, the roadside environment has improvements, with the increase of EVI. Generally, suburban areas have a higher increase of EVI than rural areas. In desert areas where environmental degradation serves, the EVI decreases a lot in the desert areas. However, road performance factors except

traffic volumes can reduce the roadside AOD, which indicates that the construction of road infrastructure is beneficial to improving the desert environment.

4.3. Sensitivity analysis

Sensitivity analysis was used to reveal the impact of the road on roadside environment change within different distances. Figure 9 shows the trend of Q values of six explanatory variables for change of roadside EVI and AOD. The order of significance of variables is not changed for all distances to roads. It shows that 1 km is the most reasonable distance to evaluate the change of roadside environment. Apart from service to industries, the road has the highest impact on EVI at 1 km for all variables. Change rates of Q values of all variables are slight, which are -1.64% to 2.44% for the change of roadside EVI and -5.81% to 1.87% for the change of roadside AOD. The road impact to the roadside EVI is most sensitive when the road buffer changes from 0.5 km to 1 km. The change rate of Q value for service to industries is 2.4% . And road impact to the roadside AOD is more sensitive from 1 km buffer to 2 km buffer, with the -1.6% change rate of Q value for service to other transportations.

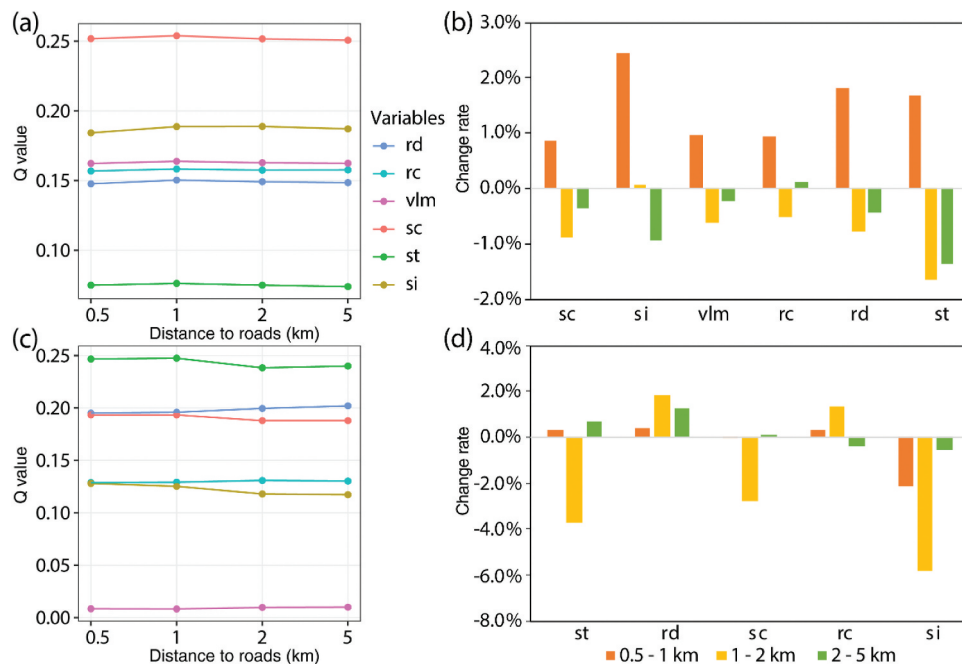


Figure 9. The Q values (left) and change rates (right) of explanatory variables at different buffers for environment variables: (a, b) change of roadside EVI; (c, d) change of roadside AOD. Roadside EVI and AOD are derived from MODIS.

5. Discussion

The positive impact of transportation infrastructure on resident income includes saving commuting time, connecting markets and raw materials, which can reduce transport costs, and providing job opportunities. However, road transportation infrastructure may damage the environment. The construction of road infrastructure and the erosion of soil in this process destroys the roadside vegetation and thus decrease the EVI. In addition, industrial facilities will come with an AOD increase. Therefore, the trade-offs between the economy and the environment must be considered when investing in road transportation infrastructure. However, past studies have been less likely to analyze and characterize this impact from a spatial perspective. Road performance indicators and environmental pollution indicators mostly use statistical data or data from monitoring stations. The spatial distribution of these data is sparse, and it is difficult to reveal spatial information. As a result, the spatial patterns of environmental and economic impacts of roads have not been explored comprehensively.

In this study, the relationship between transportation infrastructure, the environment, and the economy was explored using remote sensing data and a spatial heterogeneity model. And the trade-offs between road impacts on the local economy and the roadside environment were revealed. Findings are summaries as follows. First, road density, service to communities, and service to other transportations play the most important role in determining resident income, change of roadside EVI, and change of roadside AOD, respectively. The interaction of traffic volumes with other transport infrastructure variables can significantly affect the economy and the environment. In particular, the interaction of traffic volumes with road density can explain nearly 50% of the resident income.

Second, the environmental and economic impacts of transportation infrastructure have a spatial difference. Major cities are more dependent on transportation infrastructure for economic development but face greater environmental pressures than suburban and rural areas, with the decrease of roadside EVI and rapid increase of roadside AOD. On the contrary, the roadside environment has improved in suburban and

rural areas, with increased EVI and a lower increase of AOD than in major cities.

Third, road transportation infrastructure has a nonlinear impact on the economy and the environment, leading to the interactions of trade-offs between road impact on the economy and the impact on the local environment. The development of road infrastructure systems can enhance economic growth, but it also brings pressure to the roadside environment. In order to improve resident income and protect the environment, regional strategies are required for achieving sustainable road infrastructure development. The actions primarily include strategic road infrastructure maintenance and management (Song et al. 2018b), nation-wide and network-level strategies for sustainable infrastructure development (Song et al. 2020b), and roadside ecological and environmental protections.

In this study, results show that the effects of actions for improving both economic growth and roadside environment in suburban and rural areas are more significant than that in major cities. The primary reason is that major cities have higher environmental pressure than suburban and rural areas. This means that much more actions are required in major cities to decrease the environmental impacts of road infrastructure than in suburban and rural areas. Finally, road impact on the vegetation is most sensitive at the distance range from 0.5 km to 1 km around the road. In comparison, road impact on the AOD is most sensitive at the distance range from 1 km to 2 km around the road.

Road transportation infrastructure also has significant impacts on the landscape dynamic. First, road infrastructure has critical impacts on vegetation dynamics. According to our study, in major cities, roadside EVI is critically reduced due to dense roads, high traffic volumes, and well-constructed service facilities. Road density and traffic volumes lead to the highest EVI decrease (−1.7%). In suburban and rural areas, the roadside environment has improvements, with the increase of EVI. And suburban areas have a higher increase of EVI than in rural areas. Second, road infrastructure also has significant impacts on the fragmentation of landscape, especially in suburban and rural areas. The construction of road infrastructure can create separation and barriers, causing fragmentation of the landscapes and populations (Jaarsma and Willems 2002). Areas with high

demands of transport infrastructure have the highest fragmented landscape units (Andrea et al. 2017). To explore the environmental impact of roads, this study uses roadside EVI as the proxy of vegetation. The decrease in EVI might reveal the vegetation degradation or the increase of other land cover types. The latter reason can represent the increase of landscape fragmentation while our result can't fully prove it. Further study can use land cover data to explore the road impact on landscape fragmentation.

Road infrastructure performance is represented by six variables in this study. Road density and road connectivity are the direct indicators that characterize road infrastructure performance. Road density reflects the length of road construction, and road connectivity demonstrates the capacity of the road (Damania et al. 2018). The interaction of the two factors with traffic volume explains nearly 50% of the resident income. In major cities, road density and connectivity have the most significant positive impact on resident income. But they threaten the roadside environment, with the decrease of roadside EVI and increase of roadside AOD. In suburban areas, the road-related income is much lower than that in major cities. And the roadside environment has improved with the increase of roadside EVI.

Traffic volumes are a good indicator of economic vitality. They interact with other road variables to provide a non-linear enhancement impact on the local economy and the roadside environment. Traffic volumes also imply the generation of vehicle emissions that can cause unavoidable AOD growth. However, traffic volumes have a positive effect on roadside vegetation in suburban and rural areas, with the max increase of EVI is 0.013.

Services to facilities reflect the service performance of road infrastructure from a socio-economic perspective. High accessibility can reduce commuting time and decreases congestion, thus reducing emissions to some extent. Therefore, services to facilities lead to EVI growth in suburban and rural areas. The highest increase of EVI (2.5%) appears in areas with services to industries range from 64.4 km to 128 km. But in major cities, services to facilities would threaten the roadside environment because of more traffic volumes associated with denser facilities compared to suburban and rural areas, leading to a large amount of exhaust gas. Besides, freight transportation to ports includes many mineral fuels,

oil, gas materials, which can bring pressure to the roadside environment.

Current studies usually employed economic models or time series analysis to explore the impact of transport infrastructure on the economy and the environment, which failed to reveal the impact from the perspective of spatial (Mohmand, Wang, and Saeed 2017). Besides, enough attention has not been paid to discuss the trade-offs of the impact of the road on the economy and the environment. The main contributions of this study to road transportation research are as follows. First, road performance and roadside environmental change were evaluated by geospatial data, including POIs, population data, remote sensing data. Service to facilities was represented by accessibility using a network-based accessibility analysis method. Second, spatial trade-offs between the impact of road infrastructure on the economy and that on the roadside environment were investigated using an OPGD model. Third, the spatial difference of road impacts on the economy and the local environment was explored using mean risk values. To sum up, this study considers the heterogeneity of the spatial distribution of transport infrastructure and uses a spatial analysis model to reveal the impact of transportation infrastructure on the local economy and the environment.

There are still shortcomings in this study. First, the impact of road infrastructure performance on the economy and environment may take some time to fully manifest, but the long-time series analysis was not conducted in this study due to the limited available data. Follow-up studies should combine spatial analysis methods with a time analysis model to make a deeper analysis of the mechanism of the impact of transport infrastructure on the economy and the environment. In addition, spatial autocorrelation was not considered when conducting a spatial analysis model. Future studies should use a spatial model that combines spatial heterogeneity and spatial autocorrelation to better explain the impact of roads on the environment and economy.

6. Conclusion

This study investigates the impacts of road transportation infrastructure on the local economy and roadside environment using spatial heterogeneity methods. The spatial disparities in trade-offs have been assessed between road impacts on the

economy and environment. Significant nonlinearity and spatial disparities, such as urban-rural variations, have been identified in the trade-offs. In general, roadside EVI is decreased, and roadside AOD is critically increased in major cities together with economic growth. However, in suburban and rural areas, roadside EVI is increased, and the increase of roadside AOD is much lower than that in major cities, although the economic development is approximate with that in major cities. Therefore, this study reveals that the environmental pressure from road transportation in major cities is much higher than that in suburban and rural areas. Results show that the effects of actions for improving both economic growth and roadside environment in suburban and rural areas are more significant than that in major cities. The primary reason is that major cities have higher environmental pressure than suburban and rural areas. This means that much more actions are required in major cities to decrease the environmental impacts of road infrastructure than in suburban and rural areas. This study contributes to a deep understanding of the interaction between road infrastructure, economy, and environment and can also guide the authorities to make strategic decisions for sustainable infrastructure development.

Abbreviations

The following abbreviations are used in this manuscript:

- AOD: Aerosol Optical Depth
- EVI: Enhanced Vegetation Index
- NDVI: Normalized Difference Vegetation Index
- PD: Power of determinants
- PID: Power of interactive determinants
- POI: Point of interests
- OPGD: Optimal parameters-based geographical detector
- LGA: Local Government Area
- PWC: Population-weighted centroids
- rc: Road density
- rc: Road connectivity
- vlm: Traffic volumes
- sc: Service to communities
- si: Service to industries
- st: Service to other transportations

Acknowledgements

This research was supported by the Australian Government through the Australian Research Council's Discovery Early Career Researcher Award funding scheme (Project No.

DE170101502), and Discovery Project (Project No. DP180104026).

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Australian Research Council [DE170101502; DP180104026].

ORCID

Yongze Song  <http://orcid.org/0000-0003-3420-9622>
Peng Wu  <http://orcid.org/0000-0002-3793-0653>

References

- ABS (Australian Bureau of Statistics). 2015. *Australian Transport Economic Account: An Experimental Transport Satellite Account*. Catalogue Number 5270.0. <https://www.abs.gov.au/statistics/economy/national-accounts/australian-transport-economic-account-experimental-transport-satellite-account/latest-release>
- ABS (Australian Bureau of Statistics). 2020. *Personal Income in Australia*. Catalogue Number 6524.0.55.022. <https://www.abs.gov.au/statistics/labour/earnings-and-work-hours/personal-income-australia/2011-12-2017-18>
- Agbelie, B. R. D. K. 2014. "An Empirical Analysis of Three Econometric Frameworks for Evaluating Economic Impacts of Transportation Infrastructure Expenditures across Countries." *Transport Policy* 35: 304–310. Elsevier. doi:10.1016/j.tranpol.2014.06.009.
- Allen, C., G. Metternicht, T. Wiedmann, and M. Pedercini. 2019. "Greater Gains for Australia by Tackling All SDGs but the Last Steps Will Be the Most Challenging." *Nature Sustainability* 2 (11): 1041–1050. doi:10.1038/s41893-019-0409-9. Springer US.
- Allen, C., M. Reid, J. Thwaites, R. Glover, and T. Kestin. 2020. "Assessing National Progress and Priorities for the Sustainable Development Goals (Sdgs): Experience from Australia." *Sustainability Science* 15 (2): 521–538. doi:10.1007/s11625-019-00711-x.
- Allen, C. D., R. Ferrare, J. Szykman, J. Lewis, A. Scarino, J. Hains, and S. Burton. 2015. "Regional Characteristics of the Relationship between Columnar AOD and Surface PM_{2.5}: Application of Lidar Aerosol Extinction Profiles over Baltimore-Washington Corridor during DISCOVER-AQ." *Atmospheric Environment* 101: 338e349. Elsevier Ltd. doi:10.1016/j.atmosenv.2014.11.034.
- Allen, T., and C. Arkolakis. 2020. "The Welfare Effects of Transportation Infrastructure Improvements." *Journal of Chemical Information and Modeling* 01 (01): 1689–1699.

- Alvarez-Mendoza, C. I., A. Teodoro, and L. Ramirez-Cando. 2019. "Spatial Estimation of Surface Ozone Concentrations in Quito Ecuador with Remote Sensing Data, Air Pollution Measurements and Meteorological Variables." *Environmental Monitoring and Assessment* 191(3). Environmental Monitoring and Assessment. doi:10.1007/s10661-019-7286-6.
- Al-Yaari, A., J. P. Wigneron, W. Dorigo, A. Colliander, T. Pellarin, S. Hahn, and A. Mialon. 2019. "Assessment and Inter-Comparison of Recently Developed/Reprocessed Microwave Satellite Soil Moisture Products Using ISMN Ground-Based Measurements." *Remote Sensing of Environment* 224 (February): 289–303. doi:10.1016/j.rse.2019.02.008. Elsevier.
- Anderson, K., B. Ryan, W. Sonntag, A. Kavvada, and L. Friedl. 2017. "Earth Observation in Service of the 2030 Agenda for Sustainable Development." *Geo-Spatial Information Science* 20 (2): 77–96. doi:10.1080/10095020.2017.1333230. Taylor & Francis.
- Andrea, D. M., B. Martin, E. Ortega, A. Ledda, and V. Serra. 2017. "Landscape Fragmentation in Mediterranean Europe: A Comparative Approach." *Land Use Policy* 64: 83–94. Elsevier Ltd. doi:10.1016/j.landusepol.2017.02.028.
- Bao, Q., Z. Yuxin, W. Yuxiao, and Y. Feng. 2020. "Can Entropy Weight Method Correctly Reflect the Distinction of Water Quality Indices?." *Water Resources Management* 34 (11): 3667–3674. doi:10.1007/s11269-020-02641-1. Water Resources Management.
- Bishop-Taylor, R., M. G. Tulbure, and M. Broich. 2018. "Evaluating Static and Dynamic Landscape Connectivity Modelling Using a 25-Year Remote Sensing Time Series." *Landscape Ecology* 33 (4): 625–640. doi:10.1007/s10980-018-0624-1. Springer Netherlands.
- BITRE. 2018. *Australian Sea Freight 2015–16*. https://www.bitre.gov.au/publications/2018/asf_2015_16.aspx.
- Boullila, W., I. R. Farah, and A. Hussain. 2018. "A Novel Decision Support System for the Interpretation of Remote Sensing Big Data." *Earth Science Informatics* 11 (1): 31–45. doi:10.1007/s12145-017-0313-7. Earth Science Informatics.
- Cai, J., B. Xu, K. Kie Yan Chan, X. Zhang, B. Zhang, Z. Chen, and B. Xu. 2019. "Roles of Different Transport Modes in the Spatial Spread of the 2009 Influenza A(H1N1) Pandemic in Mainland China." *International Journal of Environmental Research and Public Health* 16(2). MDPI AG. doi:10.3390/ijerph16020222.
- Cârlan, I., D. Haase, A. Große-Stoltenberg, and I. Sandric. 2020. "Mapping Heat and Traffic Stress of Urban Park Vegetation Based on Satellite Imagery - A Comparison of Bucharest, Romania and Leipzig, Germany." *Urban Ecosystems* 23 (2): 363–377. doi:10.1007/s11252-019-00916-z. Urban Ecosystems.
- Cochran, F., J. Daniel, L. Jackson, and A. Neale. 2020. "Earth Observation-Based Ecosystem Services Indicators for National and Subnational Reporting of the Sustainable Development Goals." *Remote Sensing of Environment* 244 (February): 111796. doi:10.1016/j.rse.2020.111796. Elsevier.
- Damania, R., J. Russ, D. Wheeler, and A. F. Barra. 2018. "The Road to Growth: Measuring the Tradeoffs between Economic Growth and Ecological Destruction." *World Development* 101 (August): 351–376. doi:10.1016/j.worlddev.2017.06.001.
- Daniel(Jian), S., Z. Kaisheng, and S. Suwan. 2018. "Analyzing Spatiotemporal Traffic Line Source Emissions Based on Massive Didi Online Car-Hailing Service Data." *Transportation Research Part D: Transport and Environment* 62 (800): 699–714. doi:10.1016/j.trd.2018.04.024.
- Deljouei, A., S. Seyed Mohammad Moein, E. Abdi, M. Bernhardt-Römermann, E. L. Pascoe, and M. Marcantonio. 2018. "The Impact of Road Disturbance on Vegetation and Soil Properties in a Beech Stand, Hyrcanian Forest." *European Journal of Forest Research* 137 (6): 759–770. doi:10.1007/s10342-018-1138-8. Springer Berlin Heidelberg.
- Didan, K., A. B. Munoz, R. Solano, and A. Huete. 2015. "MODIS Vegetation Index User 'S Guide (Collection 6)." 2015 (May): 31. https://vip.arizona.edu/documents/MODIS/MODIS_VI_UsersGuide_June_2015_C6.pdf
- Ding, Y., M. Zhang, X. Qian, L. Chengren, S. Chen, and W. Wang. 2019. "Using the Geographical Detector Technique to Explore the Impact of Socioeconomic Factors on PM2.5 Concentrations in China." *Journal of Cleaner Production* 211: 1480–1490. Elsevier Ltd. doi:10.1016/j.jclepro.2018.11.159.
- Donald, S. 1968. "Two- Dimensional Interpolation Function for Irregularly- Spaced Data." *Proc 23rd Nat Conf*, New York, 517–524.
- Du, Z., X. Xiaoming, H. Zhang, W. Zhitao, and Y. Liu. 2016. "Geographical Detector-Based Identification of the Impact of Major Determinants on Aeolian Desertification Risk." *PLoS ONE* 11(6). Public Library of Science. doi:10.1371/journal.pone.0151331.
- Duarte, L., A. C. Teodoro, A. T. Monteiro, M. Cunha, and G. Hernâni. 2018. "QPhenoMetrics: An Open Source Software Application to Assess Vegetation Phenology Metrics." *Computers and Electronics in Agriculture* 148 (October 2017): 82–94. doi:10.1016/j.compag.2018.03.007. Elsevier.
- Erjia, G., R. Zhang, D. Li, X. Wei, X. Wang, and P. C. Lai. 2017. "Estimating Risks of Inapparent Avian Exposure for Human Infection: Avian Influenza Virus A (H7N9) in Zhejiang Province, China." *Scientific Reports* 7(January). Nature Publishing Group. doi:10.1038/srep40016.
- Gaughan, A. E., F. R. Stevens, C. Linard, P. Jia, and A. J. Tatem. 2013. "High Resolution Population Distribution Maps for Southeast Asia in 2010 and 2015." *PLoS ONE* 8 (2). doi:10.1371/journal.pone.0055882.
- Ge, Y. J., A. Thomasson, and R. Sui. 2011. "Remote Sensing of Soil Properties in Precision Agriculture: A Review." *Frontiers of Earth Science* 5 (3): 229–238. doi:10.1007/s11707-011-0175-0.
- Ghosh, S. P., D. Raj, and S. K. Maiti. 2020. "Risks Assessment of Heavy Metal Pollution in Roadside Soil and Vegetation of National Highway Crossing through Industrial Area."

- Environmental Processes* 7 (4): 1197–1220. doi:10.1007/s40710-020-00463-2. Environmental Processes.
- Glaeser, E. L., M. E. Kahn, and J. Rappaport. 2008. "Why Do the Poor Live in Cities? the Role of Public Transportation." *Journal of Urban Economics* 63 (1): 1–24. doi:10.1016/j.jue.2006.12.004.
- Haklay, M., and P. Weber. 2008. "Openstreetmap: User-Generated Street Maps." *IEEE Pervasive Computing* 7 (4): 12–18. doi:10.1109/MPRV.2008.80.
- Hall, N. L., S. Creamer, W. Anders, A. Slatyer, and P. S. Hill. 2020. "Water and Health Interlinkages of the Sustainable Development Goals in Remote Indigenous Australia." *Npj Clean Water* 3(4). Springer US. doi:10.1038/s41545-020-0060-z.
- Hu, Y., J. Wang, L. Xiaohong, D. Ren, and J. Zhu. 2011. "Geographical Detector-Based Risk Assessment of the under-Five Mortality in the 2008 Wenchuan Earthquake, China." *PLoS ONE* 6 (6). doi:10.1371/journal.pone.0021427.
- Im, J. 2020. "Earth Observations and Geographic Information Science for Sustainable Development Goals." *GIScience and Remote Sensing* 57 (5): 591–592. doi:10.1080/15481603.2020.1763041. Taylor & Francis.
- Jaarsma, C. F., and G. P. A. Willems. 2002. "Reducing Habitat Fragmentation by Minor Rural Roads through Traffic Calming." *Landscape and Urban Planning* 58 (2–4): 125–135. doi:10.1016/S0169-2046(01)00215-8.
- Jantunen, J., K. Saarinen, A. Valtonen, and S. Saarnio. 2006. "Grassland Vegetation along Roads Differing in Size and Traffic Density." *Annales Botanici Fennici* 43 (2): 107–117.
- Karagulian, F., C. A. Belis, C. F. C. Dora, A. M. Prüss-Ustün, S. Bonjour, H. Adair-Rohani, and M. Amann. 2015. "Contributions to Cities' Ambient Particulate Matter (PM): A Systematic Review of Local Source Contributions at Global Level." *Atmospheric Environment* 120: 475–483. Elsevier Ltd. doi:10.1016/j.atmosenv.2015.08.087.
- Khan, R., U. Haroon, M. Siddique, K. Zaman, S. U. Yousaf, A. M. Shoukry, S. Gani, A. K. Sasmoko, S. S. Hishan, and H. Saleem. 2018. "The Impact of Air Transportation, Railways Transportation, and Port Container Traffic on Energy Demand, Customs Duty, and Economic Growth: Evidence from a Panel of Low-, Middle-, and High -income Countries." *Journal of Air Transport Management* 70 (February 2017): 18–35. doi:10.1016/j.jairtraman.2018.04.013. Elsevier Ltd.
- Li, B., S. Gao, Y. Liang, Y. Kang, T. Prestby, Y. Gao, and R. Xiao. 2020a. "Estimation of Regional Economic Development Indicator from Transportation Network Analytics." *Scientific Reports* 10 (1): 1–15. doi:10.1038/s41598-020-59505-2. Springer US.
- Li, D., and Y. Liao. 2018. "Spatial Characteristics of Heavy Metals in Street Dust of Coal Railway Transportation Hubs: A Case Study in Yuanping, China." *International Journal of Environmental Research and Public Health* 15 (12). doi:10.3390/ijerph15122662.
- Li, S., D. Lyu, G. Huang, X. Zhang, F. Gao, Y. Chen, and X. Liu. 2020b. "Spatially Varying Impacts of Built Environment Factors on Rail Transit Ridership at Station Level: A Case Study in Guangzhou, China." *Journal of Transport Geography* 82 (July 2019). doi:10.1016/j.jtrangeo.2019.102631.
- Liao, Y., J. Wang, D. Wei, B. Gao, X. Liu, G. Chen, X. Song, and X. Zheng. 2017. "Using Spatial Analysis to Understand the Spatial Heterogeneity of Disability Employment in China." *Transactions in GIS* 21 (4): 647–660. doi:10.1111/tgis.12217. Blackwell Publishing Ltd.
- Ma, H., J. Zeng, N. Chen, X. Zhang, M. H. Cosh, and W. Wang. 2019. "Satellite Surface Soil Moisture from SMAP, SMOS, AMSR2 and ESA CCI: A Comprehensive Assessment Using Global Ground-Based Observations." *Remote Sensing of Environment* 231 (February): 111215. doi:10.1016/j.rse.2019.111215. Elsevier.
- Mann, D., G. Agrawal, and P. K. Joshi. 2019. "Spatio-Temporal Forest Cover Dynamics along Road Networks in the Central Himalaya." *Ecological Engineering* 127 (November 2018): 383–393. doi:10.1016/j.ecoleng.2018.12.020. Elsevier.
- Martins, V. S., A. Lyapustin, Y. Wang, D. M. Giles, A. Smirnov, I. Slutsker, and S. Korkin. 2019. "Global Validation of Columnar Water Vapor Derived from EOS MODIS-MAIAC Algorithm against the Ground-Based AERONET Observations." *Atmospheric Research* 225 (April): 181–192. doi:10.1016/j.atmosres.2019.04.005. Elsevier.
- Matsushita, B., W. Yang, J. Chen, Y. Onda, and G. Qiu. 2007. "Sensitivity of the Enhanced Vegetation Index (EVI) and Normalized Difference Vegetation Index (NDVI) to Topographic Effects: A Case Study in High-Density Cypress Forest." *Sensors* 7 (11): 2636–2651. doi:10.3390/s7112636.
- Mohmand, Y. T., A. Wang, and A. Saeed. 2017. "The Impact of Transportation Infrastructure on Economic Growth: Empirical Evidence from Pakistan." *Transportation Letters* 9 (2): 63–69. doi:10.1080/19427867.2016.1165463.
- Pathak, C., S. Chandra, G. Maurya, A. Rathore, M. O. Sarif, and R. D. Gupta. 2021. "The Effects of Land Indices on Thermal State in Surface Urban Heat Island Formation: A Case Study on Agra City in India Using Remote Sensing Data (1992–2019)." *Earth Systems and Environment* 5 (1): 135–154. doi:10.1007/s41748-020-00172-8. Springer International Publishing.
- Schultz, M., J. Voss, M. Auer, S. Carter, and A. Zipf. 2017. "Open Land Cover from OpenStreetMap and Remote Sensing." *International Journal of Applied Earth Observation and Geoinformation* 63 (May): 206–213. doi:10.1016/j.jag.2017.07.014. Elsevier.
- Shi, H., T. Shi, Z. Yang, Z. Wang, F. Han, and C. Wang. 2018. "Effect of Roads on Ecological Corridors Used for Wildlife Movement in a Natural Heritage Site." *Sustainability (Switzerland)* 10 (8): 1–24. doi:10.3390/su10082725.
- Shrestha, A., and W. Luo. 2017. "An Assessment of Groundwater Contamination in Central Valley Aquifer, California Using Geodetector Method." *Annals of GIS* (3). . doi:10.1080/19475683.2017.1346707. Taylor and Francis Ltd
- Song, Y., Y. Tan, Y. Song, P. Wu, J. C. P. Cheng, M. J. Kim, and X. Wang. 2018a. "Spatial and Temporal Variations of Spatial Population Accessibility to Public Hospitals: A Case Study of Rural–Urban Comparison." *GIScience and Remote Sensing* 55 (5): 718–744. doi:10.1080/15481603.2018.1446713. Springer US

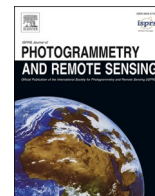
- Song, Y., G. L. Wright, P. Wu, D. Thatcher, T. McHugh, Q. Li, S. J. Li, and X. Wang. 2018b. "Segment-Based Spatial Analysis for Assessing Road Infrastructure Performance Using Monitoring Observations and Remote Sensing Data." *Remote Sensing* 10 (11). doi:10.3390/rs10111696
- Song, Y., P. Wu, D. Gilmore, and Q. Li. 2020a. "A Spatial Heterogeneity-Based Segmentation Model for Analyzing Road Deterioration Network Data in Multi-Scale Infrastructure Systems." *IEEE Transactions on Intelligent Transportation Systems* 1–11. doi:10.1109/tits.2020.3001193
- Song, Y., J. Wang, Y. Ge, and C. Xu. 2020b. "An Optimal Parameters-Based Geographical Detector Model Enhances Geographic Characteristics of Explanatory Variables for Spatial Heterogeneity Analysis: Cases with Different Types of Spatial Data." *GIScience and Remote Sensing* 57 (5): 593–610. doi:10.1080/15481603.2020.1760434. Taylor & Francis
- Song, Y., and P. Wu. 2021. "An Interactive Detector for Spatial Associations." *International Journal of Geographical Information Science*. doi:10.1080/13658816.2021.1882680
- Stevens, F. R., A. E. Gaughan, C. Linard, and A. J. Tatem. 2015a. "Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data." *PLoS ONE* 10 (2): 1–22. doi:10.1371/journal.pone.0107042.
- Umar, M., J. Xiangfeng, D. Kirikkaleli, and X. Qinghui. 2020. "COP21 Roadmap: Do Innovation, Financial Development, and Transportation Infrastructure Matter for Environmental Sustainability in China?" *Journal of Environmental Management* 271 (June): 111026. doi:10.1016/j.jenvman.2020.111026. Elsevier Ltd.
- Viana, J., J. V. Santos, R. M. Neiva, J. Souza, L. Duarte, A. C. Teodoro, and A. Freitas. 2017. "Remote Sensing in Human Health: A 10-Year Bibliometric Analysis." *Remote Sensing* 9 (12): 1–12. doi:10.3390/rs9121225.
- Wang, J., and X. Chengdong. 2017. "Geodetector: Principle and Prospective." *Acta Geographica Sinica* 72 (1): 116–134. doi:10.11821/dlxb201701010. Science Press.
- Wang, J. F., T. L. Zhang, and B. J. Fu. 2016. "A Measure of Spatial Stratified Heterogeneity." *Ecological Indicators* 67: 250–256. Elsevier Ltd. doi:10.1016/j.ecolind.2016.02.052.
- Wang, J. F., X. H. Li, G. Christakos, Y. L. Liao, T. Zhang, G. Xue, and X. Y. Zheng. 2010. "Geographical Detectors-Based Health Risk Assessment and Its Application in the Neural Tube Defects Study of the Heshun Region, China." *International Journal of Geographical Information Science* 24 (1): 107–127. doi:10.1080/13658810802443457.
- Wang, L., and L. Chen. 2018. "Analysis: The Impact of New Transportation Modes on Population Distribution in Jing-Jin-Ji Region of China." *Scientific Data* 5 (1): 1–15. doi:10.1038/sdata.2017.204. The Author(s).
- Wang, S., Y. Dexin, M. Xiaogang, and X. Xing. 2018. "Analyzing Urban Traffic Demand Distribution and the Correlation between Traffic Flow and the Built Environment Based on Detector Data and POIs." *European Transport Research Review* 10 (2). doi:10.1186/s12544-018-0325-5.
- Wang, Z., C. Fan, Q. Zhao, and S. W. Myint. 2020. "A Geographically Weighted Regression Approach to Understanding Urbanization Impacts on Urban Warming and Cooling: A Case Study of Las Vegas." *Remote Sensing* 12(2). MDPI AG. doi:10.3390/rs12020222.
- Xue ting, Y., Y. P. Fang, Q. Xiao ping, and F. B. Zhu. 2018. "Gradient Effect of Road Transportation on Economic Development in Different Geomorphic Regions." *Journal of Mountain Science* 15 (1): 181–197. doi:10.1007/s11629-017-4498-5.
- Yang, J., P. Gong, F. Rong, M. Zhang, J. Chen, S. Liang, X. Bing, J. Shi, and R. Dickinson. 2013. "The Role of Satellite Remote Sensing in Climate Change Studies." *Nature Climate Change* 3 (10): 875–883. doi:10.1038/nclimate1908.
- Yang, R., X. Qian, and H. Long. 2016. "Spatial Distribution Characteristics and Optimized Reconstruction Analysis of China's Rural Settlements during the Process of Rapid Urbanization." *Journal of Rural Studies* 47: 413–424. Elsevier Ltd. doi:10.1016/j.jrurstud.2016.05.013.
- Zhang, X., J. Nie, C. Cheng, X. Chengdong, L. Zhou, S. Shen, and Y. Pei. 2020. "Natural and Socioeconomic Factors and Their Interactive Effects on House Collapse Caused by Typhoon Mangkhut." *International Journal of Disaster Risk Science*. Beijing Normal University Press. doi:10.1007/s13753-020-00322-6.
- Zhang, Y., H. Lu, and W. Qu. 2020. "Geographical Detection of Traffic Accidents Spatial Stratified Heterogeneity and Influence Factors." *International Journal of Environmental Research and Public Health* 17(2). MDPI AG. doi:10.3390/ijerph17020572.
- Zhu, L., J. Meng, and L. Zhu. 2020. "Applying Geodetector to Disentangle the Contributions of Natural and Anthropogenic Factors to NDVI Variations in the Middle Reaches of the Heihe River Basin." *Ecological Indicators* 117 (January): 106545. doi:10.1016/j.ecolind.2020.106545. Elsevier.

A3. Identifying determinants of spatio-temporal disparities in soil moisture of the Northern Hemisphere using a geographically optimal zones-based heterogeneity model

Reference: Luo, P., Song, Y., Huang, X., Ma, H., Liu, J., Yao, Y., & Meng, L. (2022). Identifying determinants of spatio-temporal disparities in soil moisture of the Northern Hemisphere using a geographically optimal zones-based heterogeneity model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 185, 111-128.

Contents lists available at [ScienceDirect](https://www.sciencedirect.com)

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Identifying determinants of spatio-temporal disparities in soil moisture of the Northern Hemisphere using a geographically optimal zones-based heterogeneity model

Peng Luo^a, Yongze Song^{b,*}, Xin Huang^{c,d}, Hongliang Ma^c, Jin Liu^e, Yao Yao^f, Liqiu Meng^a

^a Chair of Cartography and Visual Analytics, Department of Aerospace and Geodesy, Technical University of Munich, Munich, Germany

^b School of Design and the Built Environment, Curtin University, Perth, Australia

^c State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China

^d School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China

^e State Key Laboratory of Earth Surface Processes and Resource Ecology, Faculty of Geographical Science, Beijing Normal University, Beijing, China

^f School of Geography and Information Engineering, China University of Geoscience, Wuhan, China

ARTICLE INFO

Keywords:

Spatial heterogeneity

Soil moisture

Spatial determinants

Geographically optimal zones-based heterogeneity model

Google Earth Engine (GEE)

Spatial statistics

ABSTRACT

Soil moisture is a fundamental ecological component for climate and hydrological studies. However, the distribution patterns of soil moisture are spatially heterogenous and influenced by multiple environmental factors. The knowledge is still limited in assessing the large-scale spatial heterogeneity of soil moisture in situ data modelling, in situ network design, spatial down-scaling, and remote sensing-based soil moisture retrieval. Heterogeneity models are effective in characterizing spatial disparities, but they are not capable of examining the maximum regional disparities. To address this bottleneck, the authors of this study developed a geographically optimal zones-based heterogeneity (GOZH) model. By progressively optimizing geographical zones of soil moisture and quantifying the heterogeneity among zones, GOZH may help identify individual and interactive determinants of soil moisture across a large study area. It was applied to identify spatial determinants of in situ soil moisture data collected at 653 monitoring stations in the Northern Hemisphere in unfrozen and frozen seasons from April 2015 to December 2017, with only thawed data considered in both seasons. Correspondingly, a series of variables were derived from Google Earth Engine (GEE) remote sensing data. The results demonstrated the significant regional disparities of soil moisture, and the combinations of determinants are critically different among geographical zones and between unfrozen and frozen seasons. At a global scale, the combinations of determinants can explain about 48% of the spatial pattern of soil moisture. Spatial heterogeneity of soil moisture in frozen seasons is much more complex than that in unfrozen seasons regarding geographical zones and explanatory variables. The variability of soil moisture during unfrozen seasons can be more explainable than that during frozen seasons, which was a convincing evidence for previous studies that soil moisture predictions were mostly performed during unfrozen seasons. Primary variables that determine spatial patterns of soil moisture are changed from climate variables during the unfrozen season to geographical variables during the frozen season. Results show that GOZH model can effectively explore spatial determinants of soil moisture through avoiding the underestimation of individual variables, overestimation of multiple variables, and finely divide zones. The research findings from this study provide an in-depth understanding of the spatial heterogeneity of soil moisture and can be implemented in more effective in situ sampling network design, spatial down-scaling of soil moisture, and accurate inversion of surface parameters from the satellite data of soil moisture.

1. Introduction

Soil moisture is an essential component of an ecosystem (Green et al., 2019; Li et al., 2020), and plays a fundamental role in plant growth (Lei

et al., 2018), food security (McColl et al., 2017), carbon and water cycles (Wang et al., 2017), soil productivity, and projection of the global climate change (Berg et al., 2017). Monitoring of soil moisture is required for agricultural production (Lei et al., 2018), drought

* Corresponding author.

E-mail address: yongze.song@curtin.edu.au (Y. Song).

<https://doi.org/10.1016/j.isprsjprs.2022.01.009>

Received 15 August 2021; Received in revised form 5 December 2021; Accepted 19 January 2022

0924-2716/© 2022 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). Published by Elsevier B.V. All rights reserved.

monitoring (Babaeian et al., 2018), on-farm moisture management, water resources management, hydrological simulation, and forecasting (Kumar et al., 2018). Nevertheless, the strong and complex spatial heterogeneity of soil moisture challenges large-scale and regional studies in hydrology and climate (Chaney et al., 2015; Molero et al., 2018). The spatial heterogeneity of soil moisture is closely associated with precipitation, atmospheric variability, evapotranspiration (ET), runoff etc (Quinn et al., 1995; Peters-Lidard et al., 1997; Famiglietti et al., 2008; Chaney et al., 2015). It also critically affects the in situ sampling network design (Chaney et al., 2015; Dari et al., 2019), spatial downscaling of soil moisture, hydrologic modelling, and agricultural management (Vereecken et al., 2014).

In-situ observations have been the fundamentally essential data for understanding soil moisture, climate variability, and the validation and refinements of remote sensing products (Albergel et al., 2009; Ma et al., 2019; Zappa et al., 2019; Dari et al., 2019). The in situ soil moisture observations have a number of advantages. First, the in situ data of soil moisture has high accuracy and continuous observations (Albergel et al., 2012). The in situ data are generally continuously observed by seconds at sampling sites and the accuracy can reach up to $0.05 \text{ m}^3/\text{m}^3$ (Albergel et al., 2010; Gruber et al., 2013). In addition, compared with remote sensing products with spatial resolutions between 250 m and 1 km, the in situ data are collected at precise locations of monitoring stations or GPS locations in the field (Wu and Liu, 2012). Finally, the regional, national, or global sampling networks assure the overall reliability and quality of in situ observations (Dorigo et al., 2011; Molero et al., 2018). The increasing number of open data libraries have brought about new opportunities for global data sharing and collaborative studies. For these reasons, the in situ observations provide the most essential data source for understanding soil moisture, climate variability, and the validation and refinement of remote sensing products.

The spatial variability of soil moisture is influenced by complex geographical and environmental factors, such as temperature, vegetation, topography, soil properties, depth to water table, freeze–thaw states, and scales (Brocca et al., 2010; Chaney et al., 2015; Ochsner et al., 2019; Li et al., 2021). These factors fall into four categories: (1) Climate conditions, for instance, precipitation may directly influence the water balance and the hydrological cycle (Kusangaya et al., 2016; Wang et al., 2018), and temperature may influence the energy balance principle and water circulation (Tao et al., 2021); (2) Geographical and terrain conditions, they may affect the storage and evaporation of soil moisture, and influence the direction and amount of water flow (Silva et al., 2014). The altitude may change the soil properties by influencing the environmental factor such as temperature (Niu et al., 2017); (3) Soil properties and freeze–thaw states, they are related to the forms and amount of water stored in soil. Fine-textured soils can store water more readily than coarse soils, resulting in high soil moisture. Soil texture can also affect the heat fluxes and thus soil moisture (Albergel et al., 2008; Shellito et al., 2018); (4) Surface coverage, for example, vegetation can influence the vertical drainage and ET fluxes, and closely associated with soil moisture (Baroni et al., 2013; Vereecken et al., 2014). In addition to individual variables, studies have demonstrated variables usually have interactive effects in affecting soil moisture patterns (Famiglietti and Wood, 1994; Western et al., 2004; Wilson et al., 2005; Konare et al., 2008; Chaney et al., 2015). Characteristics of spatial heterogeneity is also affected by the spatial scale (Han et al., 2018). The spatial variability of soil moisture is essentially different at the field scale (Nielsen et al., 1973; Bell et al., 1980; Vereecken et al., 2014), catchment scale (Western et al., 2004; Rosenbaum et al., 2012), regional scale (Romshoo, 2004; Zhao et al., 2013), and continental scale (Entin et al., 2000).

Given the considerable differences among regional environmental impacts on soil moisture, effective and reliable geographical zones are critically important for regional soil moisture inversions from remote sensing data, downscaling with the supports of local terrain and environmental variables (Zhuo et al., 2020), and network design (Vereecken et al., 2014). For example, the valid reference of regional difference of

the soil moisture determinants is increasingly needed at a global scale when calibrating the ground roughness parameterization scheme with ground observation data (Verhoest et al., 2008). The limited knowledge in the regional disparity of soil moisture and its controls have been a challenge for the interpretability and transferability of the parameters. In addition, the geographical zones considering the spatial heterogeneity of soil moisture can support the network design (Zhuo et al., 2020). The network can capture spatial variability of soil moisture at the lowest possible cost by improving the representation of the soil moisture samples (Vereecken et al., 2014; Chaney et al., 2015).

A wide range of methods have been developed to understand the spatial heterogeneity of soil moisture. The commonly use methods can be classified into three categories, geostatistical analysis, wavelet analysis, and empirical orthogonal function (EOF) (Vereecken et al., 2014). Geostatistical models are effectively applied to identify static mapping patterns in soil properties (Ochsner et al., 2019). The spatial pattern of soil moisture at the field scale determined by multiple factors was observed through geostatistical analysis (Brocca et al., 2010). Wavelet analysis was originally used to analyze time series and has been applied to characterize the spatial variability of soil data patterns (Song et al., 2021; Vereecken et al., 2014). For example, the spatial pattern of soil moisture and temperature in the southern interior of British Columbia was characterized based on wavelet analysis (Redding, 2003). The difference of spatial scales in soil moisture variability was revealed using the wavelet analysis (Das and Mohanty, 2008). Empirical orthogonal function (EOF) methods were developed in terms of spatial modes and signal processing of soil moisture data (Wang et al., 2017). For instance, studies based on the EOF methods demonstrate that soil characteristics and topography were the two most critical factors to soil moisture (Perry and Niemann, 2007), and soil texture explained 61% of the variation in soil moisture (Jawson and Niemann, 2007).

Spatial stratified heterogeneity (SSH) models are effective approaches to investigate determinants of spatial variability of geographical variables (Wang et al., 2016). The basic assumption of SSH models is to compare the zonal spatial distribution patterns of dependent and independent variables. The zones are determined by categories of categorical variables or the spatial discretization of continuous explanatory variables (Song et al., 2020; Wang et al., 2010). As such, the spatial discretization is essential for identifying spatial determinants, and the process of spatial discretization is presented in the next paragraph. The power of determinants (PD) is calculated as a ratio between the sum of zonal variance and the variance of data across the whole space. This means that a higher PD value is associated with higher zonal variance. The commonly used SSH models include Geodetector (Song et al., 2018; Wang et al., 2010, 2016), optimal parameter-based geographical detector (OPGD) (Song et al., 2020; Luo et al., 2021), interactive detector of spatial associations (Song and Wu, 2021), etc. The SSH models have been increasingly implemented to characterize the spatial variability of soil properties. For example, the spatial difference of tillage factors of the China soil loss equation was characterized using the SSH model (Chen, 2021). The driving forces of soil erosion were explored using the GD model (Liang and Fang, 2021). The spatiotemporal variability of soil organic matter was also revealed based on the heterogeneity using the GD model (Hu et al., 2021). In addition, some soil properties, such as soil organic carbon, were mapped using the GD-based kriging model (Liu et al., 2021). Overall, existing research demonstrated the effectiveness and viability of using the spatial stratified heterogeneity model to reveal the variability of soil variables.

However, regarding the complex spatial heterogeneity of in situ soil moisture in large regions, there are still difficulties in addressing following issues using current SSH models. First, spatial discretization is an essential step to identify geographical zones based on spatial patterns of explanatory variables (Song and Wu, 2021). In current studies, the general procedure of spatial discretization is performed using a two-step approach. The individual geographical variables are first discretized using supervised or unsupervised approaches, such as equal, quantile,

standard deviation, and geometric breaks, to determine spatial zones based on an individual variable, and then combine the spatial zones through a spatial overlay (Cang and Luo, 2018; Song et al., 2020). In this method, distribution characteristics of the response variables are not fully explained in the discretization process, leading to the incomplete exploration of the influence of explanatory variables on the response variable. Thus, it is necessary to identify the geographically optimal zones which can maximize the difference among zones and minimize the similarities within zones. In addition, the reliability of estimating the power of interactive determinants needs to be improved due to the massive finely divided zones from the spatial overlay of zone layers of multiple explanatory variables (Song and Wu, 2021). In most of the previous studies, the power of interactive determinants is only estimated for the interaction of only two or three explanatory variables as for the problem of massive finely divided zones. Therefore, it is essential to develop reliable models to identify geographical optimal zones and more accurately estimate the power of interactive determinants of spatial heterogeneity of the soil moisture in large regions.

In this study, we developed a geographically optimal zones-based heterogeneity (GOZH) model to characterize the spatial heterogeneity and examine determinants of large-scale soil moisture. In the GOZH model, an optimal power of determinant (OPD) indicator was developed to reveal the contributions of variables on spatial patterns of soil moisture, where the spatial discretization was conducted heuristically in a step-wise process. The GOZH model was used to identify the geographically optimal zones during the unfrozen and frozen season and estimate the determinants of spatio-temporal disparities in soil moisture of the Northern Hemisphere. Soil moisture in situ data were collected at 653 monitoring stations in the Northern Hemisphere from April 2015 to December 2017 to present the soil moisture with high accuracy and in precise locations. Only soil data at thawed status were included to ensure the modelling reliability. Correspondingly, remote sensing-based explanatory variables were derived from Google Earth Engine (GEE), and classified into four categories, geography, climate, soil, environment ecology. First, impacts of individual variables on soil moisture and temporal variations during 33 months were characterised. Second, the geographically optimal zones of seasonal soil moisture were identified using the GOZH model. Third, the determinants of spatial patterns were demonstrated during two seasons according to the geographically optimal zones derived in the last step. Finally, the performance of GOZH model was evaluated and compared with the OPGD model.

The remainder of this paper is structured as follows: Section 2 introduces the in situ soil moisture data and explanatory variables used in this study. Section 3 describes the objective, definition, and derivation of the GOZH model. Section 4 covers the methodologies to explore the soil moisture variability in the Northern Hemisphere using the GOZH model. Section 5 presents results of this study, including impacts of individual variables and temporal variations, geographically optimal zones, and determinants of spatial disparities and seasonal effects. Findings and research contributions are discussed in Section 6, and the study is concluded in Section 7.

2. Data

2.1. In-situ soil moisture data

In this study, monthly in situ soil moisture data in 762 observation locations from 653 monitoring stations across the Northern Hemisphere were selected to reveal the heterogeneity and determinants of soil moisture (Table 1). All stations belong to 12 networks in the International Soil Moisture Network (ISMN) (Dorigo et al., 2011; Dorigo et al., 2021). ISMN is a widely used soil moisture network that collects soil moisture and soil temperature data sets from global networks, including 1,400 stations from 40 global networks of soil monitoring (Dorigo et al., 2015; Ma et al., 2019; Ma et al., 2021). As a data hosting facility of soil moisture data, ISMN has been widely used for validations of satellite-

Table 1

A brief description and sources of soil moisture in situ data used in this study.

Network name	Country	Number of stations	Depth (cm)	Reference
USCRN	America	97	5	(Bell et al., 2013)
SNOTEL	America	208	5	https://www.wcc.nrcs.usda.gov/
SCAN	America	157	5	(Schaefer et al., 2007)
CTP_SMTMN	China	53	0–5	(Yang et al., 2013)
RISMA	Canada	14	0–5	(McNairn et al., 2014)
HOBE	Denmark	28	0–5	(Jensen and Illangasekare, 2011)
FMI	Finland	19	5	(Zeng et al., 2016)
SMOSMANIA	France	15	5	(Albergel et al., 2008)
TERENO	Germany	4	5	(Zacharias et al., 2011)
BIEBRZA-S-1	Poland	18	5	http://www.igik.edu.pl/en
REMEDIHUS	Spain	20	0–5	(Martínez-Fernández and Ceballos, 2005)
RSMN	Romania	20	0–5	(Ma et al., 2019)

derived soil moisture products (Ma et al., 2019; Dorigo et al., 2021; Ma et al., 2021). In each station, soil moisture data ranging from April 2015 to December 2017 were collected and analyzed to characterize the spatial and temporal patterns of soil moisture in the Northern Hemisphere.

2.2. Explanatory variables

Four categories of explanatory variables have been collected to explain the spatial disparities of soil moisture. They include geographical, climate, soil, and environmental variables derived from remote sensing data (Table 2). All the remote sensing data were derived and processed using the Google Earth Engine (GEE). Climate and environmental variables, with the temporal resolution from 8 days to one month, were collected from April 2015 to December 2017, consistent with the in situ soil moisture data.

(A) Geographical variables

Four terrain explanatory variables, including elevation, slope, aspect, and hill shade, were included in this study to demonstrate the local geographical conditions. The terrain variables were derived from the Shuttle Radar Topography Mission (SRTM) data. The SRTM provides the digital elevation model (DEM) data with the resolution of about 30 m (Elkhrachy, 2018). Slope, aspect, and hill shade variables were

Table 2

Explanatory variables of the spatial disparities of soil moisture.

Category	Variable	Product	Temporal resolution
Geography	Elevation	SRTM DEM	-
	Slope	SRTM DEM	-
	Aspect	SRTM DEM	-
	Hill shade	SRTM DEM	-
Climate	Precipitation	GPM	Monthly
	Temperature	MOD11	8 days
Soil	Soil texture	OpenLandMap	-
	Soil pH	OpenLandMap	-
	Soil bulk density	OpenLandMap	-
Environment	Normalized difference vegetation index (NDVI)	MOD13Q1	16 days
	Enhanced Vegetation Index (EVI)	MOD13Q1	16 days
	Leaf Area Index (LAI)	MOD15A2H	8 days
	Evapotranspiration	MOD16A2	8 days

calculated based on the DEM data using GEE spatial analysis, where local gradients were computed using the 4-connected neighbors of each pixel for the calculation of slope and aspect.

(B) Climate variables

Precipitation and temperature are two essential climate controls on soil moisture at a large spatial scale. In this study, monthly precipitation data was derived from the Global Precipitation Measurement (GPM) IMERG Final Precipitation L3 1 month 0.1 degree x 0.1 degree V06 (Joyce and Xie, 2011; Hou et al., 2014). The GPM is an international satellite mission to provide precipitation data at a 0.1-degree resolution. The temperature data were taken from the land surface temperature (LST) product of the 8-days MOD11 composition with a spatial resolution of 1.2 km (Hashimoto et al., 2008). LST has been used as an effective data source for assessing soil conditions and forecasting the soil moisture (Holzman et al., 2014; Jiang and Weng, 2017).

(C) Soil properties

The soil properties used in this study include soil texture, soil pH, and soil bulk density extracted from the Soil Moisture Active Passive (SMAP) product (Entekhabi et al., 2010). The SMAP provides a series of soil properties data at different depths between 10 cm and 200 cm, and a resolution of 250 m. Corresponding to the soil moisture data that were collected at depths of 0 to 5 cm from soil monitoring networks (Table 1), soil texture, pH and bulk density data were collected at a 10 cm depth to represent soil properties related controls of soil moisture.

(D) Environmental variables

Local environmental and ecological conditions surrounding soil moisture monitoring stations were characterized using Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), Evapotranspiration (ET), and Leaf Area Index (LAI) data driven from MODIS products using GEE. The NDVI and EVI data were extracted from the 16-days Terra MODIS products (MOD13Q1) with a spatial resolution of 250 m (Didan et al., 2015). The ET data were collected from the global 8-day MOD16A2 product with a 1 km spatial resolution (Mu et al., 2013). The ET variable is used to present the water cycle of the Earth's climate system, especially the evaporation and transpiration processes that are critically associated with soil moisture (Purdy et al., 2018). LAI data, used as an ecology indicator to explore the soil moisture variability, was extracted from the MOD15A2H product, which was a 8-days composite data with a 500 m resolution. LAI is defined as the one-sided green leaf area per unit ground area in broadleaf canopies and as one-half the total needle surface area per unit ground area in coniferous canopies. LAI is an essential indicator of vegetation structure for revealing the interaction between soil and vegetation (Fang et al., 2019). The interaction of LAI and soil moisture has a significant impact on drought, vegetation growth, vegetation senescence, and drought forest (Sawada, 2018; Liu et al., 2017).

The explanatory variables are derived from the pixels at the location of the soil moisture monitoring stations. In this study, spatial heterogeneity of soil moisture at stations in the Northern Hemisphere is much higher than that of data within grids of explanatory variables, e.g. 90 m or 250 m. Therefore, spatial analysis in the study will not be affected by the scale effects of explanatory variables derived from remote sensing or grid data. Similar processing of deriving explanatory variables of soil moisture from grid data can be found in (Peng et al., 2015; Qu et al., 2021).

2.3. SMAP freeze/thaw product

To ensure the reliability of soil moisture analysis, only the in situ data of soil moisture at thaw-status landscape were used in the study. To

select the monthly in situ data at the unfrozen situation, the SMAP L3 Freeze/Thaw product (SPL3FTP) with a 36 km resolution in the Northern Hemisphere were collected (Xu et al., 2018). The SMAP is a NASA satellite mission launched in 2015 to monitor the surface (about 5 cm) global soil moisture and landscape freeze/thaw status (Entekhabi et al., 2010; Al-Yaari et al., 2019). Missing data of the SMAP product was filled through the comparison of data at neighbouring locations and periods. The thaw or freeze status of the landscape at the soil moisture stations were derived through spatial overlay.

3. Geographically optimal zones-based heterogeneity (GOZH) model

3.1. Power of determinants (PD) of spatial stratified heterogeneity (SSH) model

As introduced above, in SSH models, a higher PD value of an explanatory variable indicates that the spatial distribution pattern of this variable tends to be more similar to the spatial pattern of response variable, i.e., soil moisture in this study. The process of estimating PD values of explanatory variables generally includes three steps. First, continuous explanatory variables should be converted to stratified variables using spatial discretization methods. The stratified variables can determine a series of geographical zones of soil moisture. Second, if multiple variables are used to identify the interactive impacts on soil moisture, geographical zones determined by explanatory variables need to be overlapped to generate a new layer of geographical zones, which contain geoinformation of all the variables. Finally, the PD value for the comparison of spatial patterns between response variable and explanatory variables are calculated as a ratio of the variance of soil moisture within geographical zones determined by one or multiple explanatory variables and the variance across the whole study area. The PD is computed as:

$$PD = 1 - \frac{SSW}{SST} = 1 - \frac{\sum_{z=1}^h N_z \sigma_z^2}{N \sigma^2} \quad (1)$$

where SSW is the Sum of Squares Within geographical zones determined by explanatory variables, SST is the Sum of Squares Total of soil moisture in the whole study area, N_z and σ_z are the number and standard deviation of soil moisture within geographical zone z ($z = 1, \dots, h$), and N and σ are the number and standard deviation of soil moisture across the study area. PD value ranges from 0 to 1, where a high PD value indicates a high spatial association between response variable and the explanatory variable.

From this equation and recent studies, we can find that the PD value is sensitive to the geographical zones determined by the spatial discretization of explanatory variables. As such, a more effective and reliable spatial discretization approach is required to maximize the variance values among zones and minimize variance within zones. In addition, as explained in the introduction section, reliable geographical zones are also essential for regional soil moisture inversions from remote sensing data and downscaling with the supports of local terrain and environmental variables.

3.2. PD of GOZH model

In this study, we define the PD as a function of explanatory variables and geographical zones, which are determined by stratified variables from certain spatial discretization processes:

$$\gamma(X, D) = 1 - \frac{SSW_{X,D}}{SST} \quad (2)$$

where X is one or multiple explanatory variables, D is the stratified variable for describing geographical zones, and $SSW_{X,D}$ is the sum of

squares within geographical zones that are recorded as D and determined by explanatory variable X .

In GOZH model, the optimal PD (OPD) value can demonstrate the maximum explanatory power of variables in terms of geographically optimal zones. As such, the OPD of explanatory variables, expressed as Ω value, is the maximum value of the PD function γ :

$$\Omega = \max(\gamma) = 1 - \frac{\min(SSW_{X,D})}{SST} \tag{3}$$

The geographically optimal zones have the minimum intra-area variance and the maximum inter-area variance. To calculate the Ω value, an optimization process is performed as:

$$\min(SSW_{X,D}) = \min \left\{ \sum_{z=1}^h \sum_{j=1}^{N_z} (y_{z,j} - \bar{c}_z)^2 \right\} \tag{4}$$

where $y_{z,j}$ and \bar{c}_z are the j th observation and mean values of soil moisture in zone z , respectively.

This equation is a nondeterministic polynomial-time complete (NP-complete) problem, which is difficult to derive a global optimum. To solve this equation, a step-wise spatial discretization of soil moisture is performed using a heuristic method with spatial explanatory variables. First, all possible two-zone solutions of spatial discretization are derived for explanatory variables, and the optimal one is selected as the cutoff point according to the squared error minimization criterion. An iteration process is performed for each variable X_k to identify the optimal cutoff point s , and the parameters in the iteration can be presented as (k, s) .

Accordingly, the input space is divided into two regions. Second, the iteration process is performed for multiple variables. The k th explanatory variable X_k and its fetching value s_k are used as cut-off variables and cut-off points, respectively. Two regions in each iteration are defined as $R_1(k, s) = \{x | x^{(k)} \leq s\}$ and $R_2(k, s) = \{x | x^{(k)} > s\}$. In each split, the variable that allows the maximum explanation of the variance of the dependent variable is selected. Thus, the optimization process is converted to a process to identify the optimal variable X_k and the cutoff point s of variable X_k , which can be expressed as:

$$\min_{k,s} \left\{ \sum_{x_i \in R_1(k,s)} (y_i - \bar{d}_1)^2 + \sum_{x_i \in R_2(k,s)} (y_i - \bar{d}_2)^2 \right\} \tag{5}$$

where \bar{d}_1 and \bar{d}_2 are the average values of soil moisture in group R_1 and R_2 , respectively. Thus, the above discretization process is repeated within each group until the data volume of the group less than a certain number, which is called minsplit. During the step-wise spatial discretization, when the data volume in one group is less than the minsplit, this group would not be subdivided further and automatically becomes a final spatial zone. This process is similar to the classification and regression tree (CART) algorithm (Breiman et al., 2017). The whole spatial discretization can be visualized as the binary tree structure.

Fig. 1 shows an example of the spatial discretization process of the GOZH model. In this example, the response variable is D and explanatory variables include A , B , and C . To conduct the step-wise spatial discretization, explanatory variables are processed one by one. For each variable, a series of cutoff points are selected to split the study area into

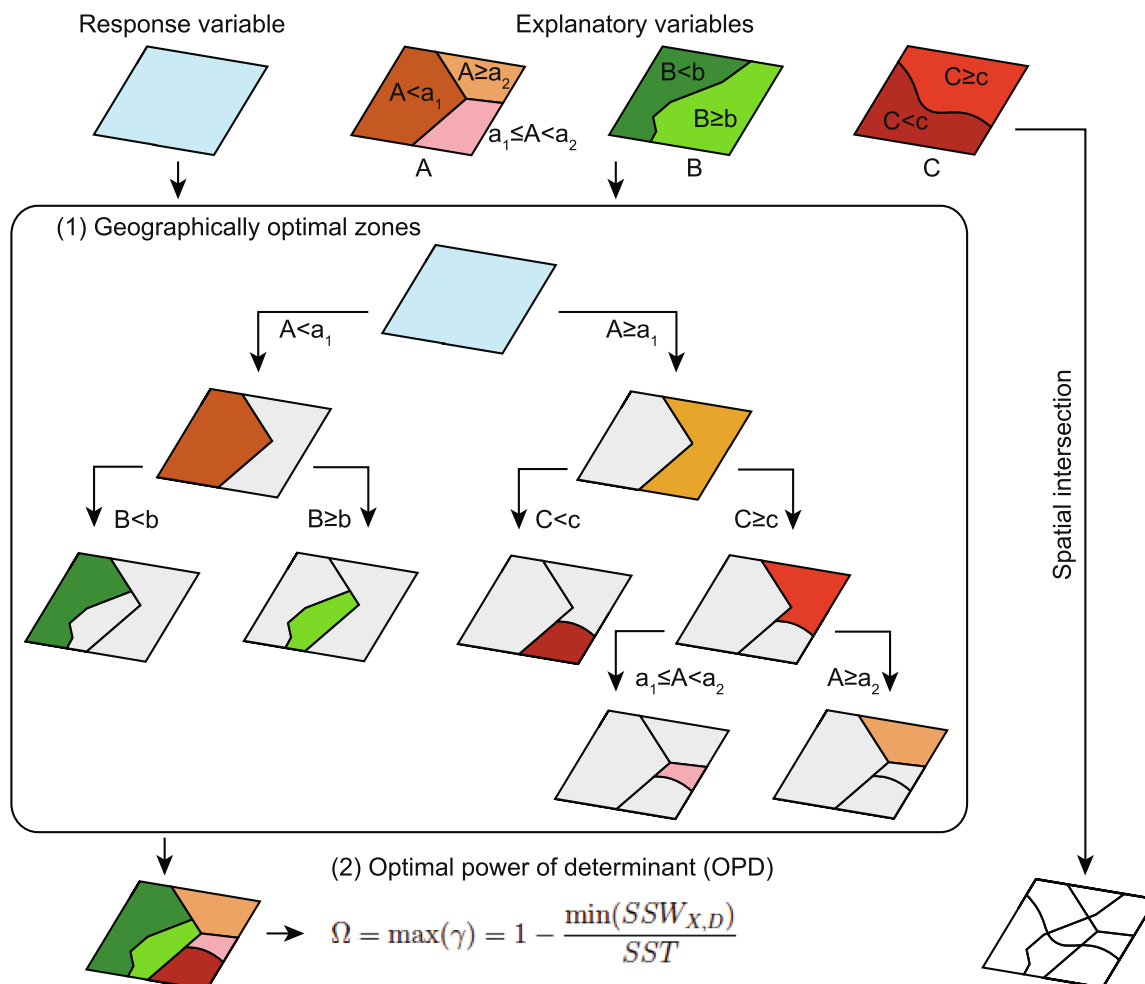


Fig. 1. Process and principle of the geographically optimal zones-based heterogeneity (GOZH) model.

two zones, and the *SSW* of soil moisture is calculated. Among all possible two-zone solutions, the one with the highest *SSW* is regarded as the optimal spatial discretization result, i.e., the optimal geographical zones, in this step. Repeat this process for each variable and data within zones, and finally, the optimal geographical zones with the highest overall *SSW* are regarded as the optimal discretization variable.

Compared with *PD* values in *SSH* models, the Ω values of *GOZH* model can identify the optimal geographical zones of data and demonstrate the maximum *PD* of explanatory variables. The *GOZH* model also can more effectively reveal the interaction effects of variables compared with *SSH* models.

4. GOZH-based spatiotemporal determinants and heterogeneity of soil moisture

Fig. 2 shows a flowchart of the *GOZH*-based spatio-temporal determinants and heterogeneity analysis of soil moisture in the Northern Hemisphere. The methods include five steps. The first step was the data pre-processing of in situ soil moisture data and explanatory variables data. Second, the monthly Ω values of individual explanatory variables were calculated to assess contributions of variables to spatial patterns of soil moisture. Third, the geographically optimal zones of soil moisture in unfrozen and frozen seasons were identified respectively using the *GOZH* model. The fourth step was to calculate determinants of spatial heterogeneity in soil moisture in unfrozen and frozen seasons with the support of the geographically optimal zones identified in the previous step. Finally, model validation was performed to assess the reliability and effectiveness of the model.

4.1. Data pre-processing

The in situ soil moisture data and explanatory variables data were first processed before modelling. The data pre-processing consists of following three parts. First, monthly in situ soil moisture at thaw landscape were selected using the corresponding *SMAP Freeze/Thaw* data. Second, the temporal periods soil moisture data were divided into unfrozen and frozen seasons to characterize respective spatial and temporal variation patterns of soil moisture. In this study, since most of the stations are located in the mid-latitudes of the Northern Hemisphere, April to September was regarded as unfrozen seasons and remaining months were frozen seasons. Finally, explanatory variables were processed to corresponding locations and time periods of soil moisture monitoring stations. For instance, the 8-day composite products *LST*, *LAI*, and *Evapotranspiration*, and the 16-day composite data product *NDVI* and *EVI* were processed to monthly data using *GEE*. A small amount of missing data in a few variables was interpolated using an inverse distance weighting (*IDW*) spatial interpolation approach.

4.2. Impacts of individual variables and their temporal variations

In the study, the *GOZH* model is first implemented in investigating impacts of individual explanatory variables on spatial patterns of soil moisture in each month from April 2015 to December 2017. In the monthly *GOZH* model, the geographical optimal zones determined by an individual variable were identified through an iteration process to maximize the variance among zones and minimize variance within zones. The *minsplit* was selected as 30 according to the data volume (i.e.

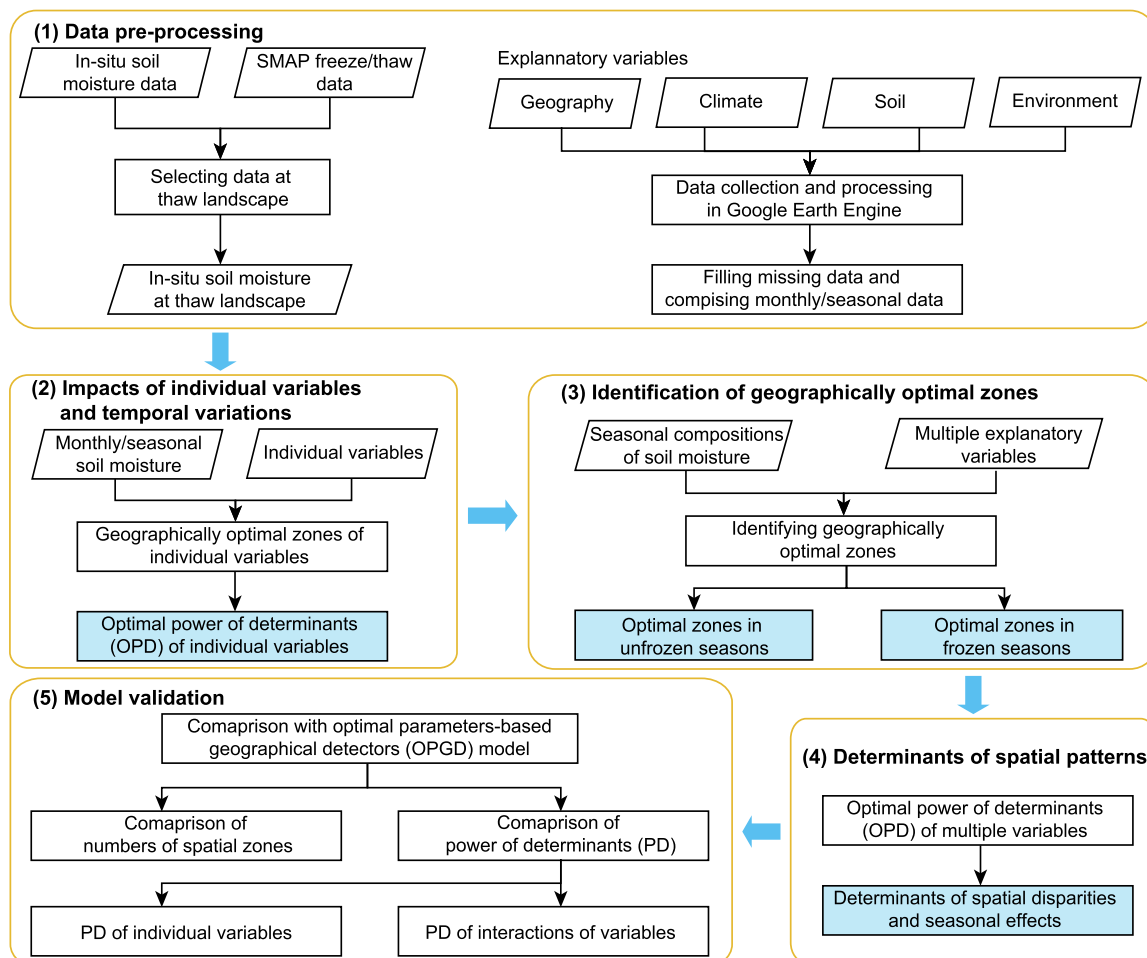


Fig. 2. Schematic overview of *GOZH*-based identification of determinants for spatiotemporal heterogeneity of soil moisture in the Northern Hemisphere.

653 stations) to avoid finely divided zones. During the spatial discretization process, the group with less than 30 stations would not be further subdivided and regarded as final spatial zones. Then, an Ω value is calculated to present the PD of this variable on soil moisture. Since explanatory variables were classified into four aforementioned categories, Ω value were computed for each variable in the four categories and in each month to indicate the temporal variations of impacts of individual variables on spatial patterns of soil moisture.

4.3. Identification of geographically optimal zones

Reliable and effective geographical zones of soil moisture are essential for parameter inversions from remote sensing data and an accurate downscaling. In the study, geographically optimal zones of soil moisture based on multiple variables were identified using the GOZH model, indicating the highest homogeneity within zones and the highest heterogeneity between zones. The optimal zones were then applied to assess the overall impacts of multiple variables on spatial patterns of soil moisture in the next step.

The identification of geographically optimal zones for unfrozen and frozen seasons takes place in three stages. First, values of monthly soil moisture and explanatory variables were merged for each season. Second, the optimal interaction was explored for each season, and the corresponding stratified variable from the spatial discretization was used to identify the geographically optimal zones. According to the stratified variable, soil monitoring stations were grouped into zones from the optimal interaction. Finally, geographical, climate, soil, and environmental characteristics at the locations of soil monitoring stations were summarized and analyzed according to the geographically optimal zones to reveal the regional spatial variability of soil moisture at a global scale.

4.4. Determinants of spatial disparities and seasonal effects

This step aims at quantifying the overall Ω value of the spatial interaction of explanatory variables on spatial patterns of soil moisture in unfrozen and frozen seasons. To assess the Ω value, an optimal interaction variable was created using the geographically optimal zones identified in the previous step. The optimal interaction variable was a categorical variable that involved the inter-dependencies of different explanatory variables and could control the spatial variability of soil moisture. Assuming the total of number of variables was n , the total number of possible spatial interactions (i.e., combinations) of variables was M ($M = 2^n - n - 1$). The optimal interaction variable demonstrated the highest Ω value among all potential spatial interactions of variables.

In addition to the overall Ω value of the spatial interaction of multiple explanatory variables, contributions of each variable within the overall Ω value was calculated using a variable removal method. The reduction of the Ω value due to the removal of this variable was calculated by removing each explanatory variable one by one in the optimal combination. The percentage of the Ω reduction of a given variable among the sum of the Ω reduction of all variables indicated the relative importance of this variable. Finally, the contribution of a given variable to spatial patterns of soil moisture was defined as the overall Ω value multiplied by its relative importance. This variable removal method has been widely applied in identifying contributions of variables within a total contribution in nonlinear models, such as generalized additive models (Song et al., 2015).

4.5. Model evaluation

To evaluate the effectiveness and reliability of the proposed GOZH model, a set of indicators were developed for comparing model performance between GOZH and the commonly used OPGD model. The indicators include PD values of individual variables, PD values of interactions of variables, and the number of geographical zones for

examining interactive effects of variables. The OPGD is an improved geographical detector model, which can be used to estimate PD values of both individual variables and interactions of variables by optimizing the spatial discretization process using unsupervised or supervised approaches (Song et al., 2020). In the OPGD model in this study, the discretization method is quantile breaks and the optional numbers of discretization are consecutive integers from 3 to 22. For each optional number of discretization, PD values were computed for all explanatory variables. Then, a local estimated scatter plot smoothing (LOESS) function was applied to model the trend of the 75% quantile values of PD values and calculate the change rates of the trend, where the span for fitting the LOESS function was 0.75 (Luo et al., 2021; Song and Wu, 2021). Finally, the break number enabled the change rate lower than 5% is selected as the optimal break number. All these parameters are selected based on the parameter selection approaches in previous studies (Song and Wu, 2021). The OPGD model was performed using the "GD" package in R (Song et al., 2020).

5. Results

5.1. Spatial and temporal patterns of soil moisture

Fig. 3 shows spatial distributions of monthly mean in situ soil moisture in the Northern Hemisphere in unfrozen (April–September) and frozen (October–March) seasons from 2015 to 2017. In general, soil moisture monitoring stations used in the study are densely distributed in North America (635 observations), and other stations are distributed in Europe (181 observations) and in China (53 observations). Along the longitude, locations of soil moisture monitoring stations can be classified into four areas as illustrated in Figs. 3 b and d. The spatial disparities of in situ soil moisture in Europe tend to be higher than those in other regions. In addition, the small figure on the right side of the Fig. 3 b shows the seasonal effects of monthly soil moisture at both all monitoring stations and stations at the thawed landscape. The seasonal effects show that the monthly mean soil moisture generally peaks in March and has the lowest values in July. The soil moisture in thawed locations tends to be higher than that in frozen locations. For instance, in March 2017, the mean soil moisture at all stations was 0.27, but at thawed locations was 0.29.

5.2. Impacts of individual variables and their temporal variations

The GOZH model first identified the primary variables of soil moisture. Fig. 4 shows the Ω values of different categories of explanatory variables on spatial patterns of soil moisture and their temporal variations in the study period. The monthly variations of Ω values indicate that spatial associations between patterns of soil moisture and explanatory variables have similar temporal trends to the spatial variability of soil moisture, which is marked with black lines in Fig. 4 a - d. This consistent trends demonstrate the effectiveness of the Ω values in examining spatial disparities of soil moisture. Fig. 4 e shows monthly average Ω values from April 2015 to December 2017. Among the four categories of variables, climate variables have the highest spatial associations with soil moisture, followed by geographical and environmental variables. For instance, from the perspective of individual variables, precipitation, elevation, and temperature have the highest Ω values among 13 variables and across the 33 months. The maximum Ω value is the impact of precipitation (58%) in November 2016. In this month, elevation and temperature can explain 55% and 46% of the spatial variability of soil moisture, respectively. Compared with climate, geographical, and environmental variables, soil property variables tend to have lower spatial associations with soil moisture, where soil texture has the lowest Ω values, ranging from 0% to 12%.

From the perspective of monthly variations, in the transitional months from frozen to unfrozen seasons, i.e., March and April, spatial associations between patterns of soil moisture and explanatory variables

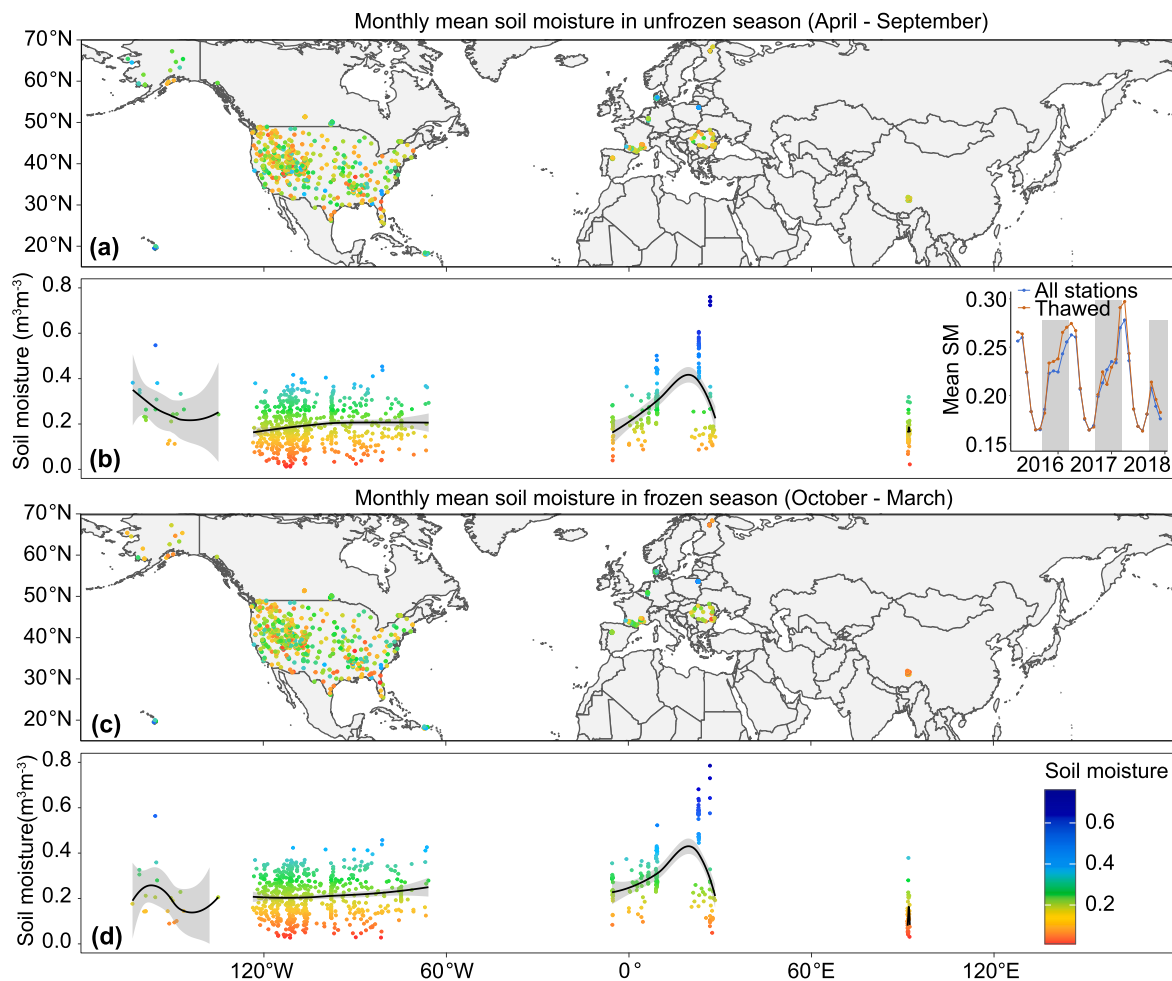


Fig. 3. Spatial distributions of monthly mean in situ soil moisture in the Northern Hemisphere in unfrozen (a and b) and frozen (c and d) seasons. In the small figure inside B, white background shows unfrozen seasons, and gray background shows frozen season.

generally have the highest Ω values. This means that spatial patterns of soil moisture in frozen-unfrozen transitional months can be more explained by geographical, climate, soil, and environmental variables. For instance, impacts of climate variables on patterns of soil moisture have been high in March ranging from 42% to 48% and with an average Ω value 47%. In March, the average Ω values of geographical, environmental, and soil variables are 24%, 22%, and 9%, respectively, which all are the highest monthly average Ω values in each category. We also can find that the least interpretable period of spatial patterns of soil moisture is the middle frozen seasons, i.e. December and January.

From the perspective of seasonal variations, results in Fig. 5 also reveal that the spatial pattern of soil moisture is more interpretable during the unfrozen season compared with that during the frozen season. The spatial associations in unfrozen and frozen seasons can be explained in a number of aspects. First, precipitation, elevation, and temperature have been the variables with the highest spatial associations with soil moisture. Ω values of precipitation, elevation, and temperature are 37.1%, 35.9%, and 32.1% in the unfrozen season, respectively, and 31.3%, 37.3%, and 34.5% in the frozen season, respectively. Ω values of environmental variables, including NDVI, LAI, EI and EVI, are lower than those of climate variables and elevation. Their contributions to spatial patterns of soil moisture are 15.8%-28.4% during the unfrozen season, and 9.5%-23.9% during the frozen season. This means that environmental variables also make important contributions to spatial variability of soil moisture. Soil property variables have the lowest Ω values in both unfrozen and frozen seasons. In addition, Ω values of most variables in the four categories have been reduced from

unfrozen to frozen seasons. The average Ω value of individual variables during the unfrozen season (20.0%) is 12.4% higher than that during the frozen season (17.8%). Ω values of precipitation, environmental variables, hill shade, and soil property variables during the unfrozen season are 5.8%, 3.4%-6.4%, 1.3%, and 0.4%-2.5% lower than that during the frozen season. Third, different from most variables, Ω values of temperature and geographical variables, including elevation and aspect, are increased from unfrozen to frozen seasons. Compared with unfrozen season, the Ω value of temperature, elevation, and aspect during the unfrozen season are 2.4%, 1.5%, and 1.3% lower than that during the frozen season. Finally, the above results also indicate that the spatial variability of soil moisture is complex and it is difficult to be explained by individual variables. The maximum interpretability of spatial patterns of soil moisture is only around 37% using individual variables.

5.3. Geographically optimal zones

5.3.1. Unfrozen seasons

Fig. 6 shows the geographically optimal zones of soil moisture during the unfrozen season identified using the GOZH model. The geographically optimal zones during the unfrozen season were identified using four explanatory variables, including precipitation, NDVI, temperature, and soil pH, and included nine zones. Fig. 6 b shows that precipitation was the primary variable that controlled spatial patterns of soil moisture during the unfrozen season. According to parameters of precipitation in the top two layers, the nine zones can be classified into three groups: the first group (precipitation < 0.082 mm/hr) contained zone A, the second

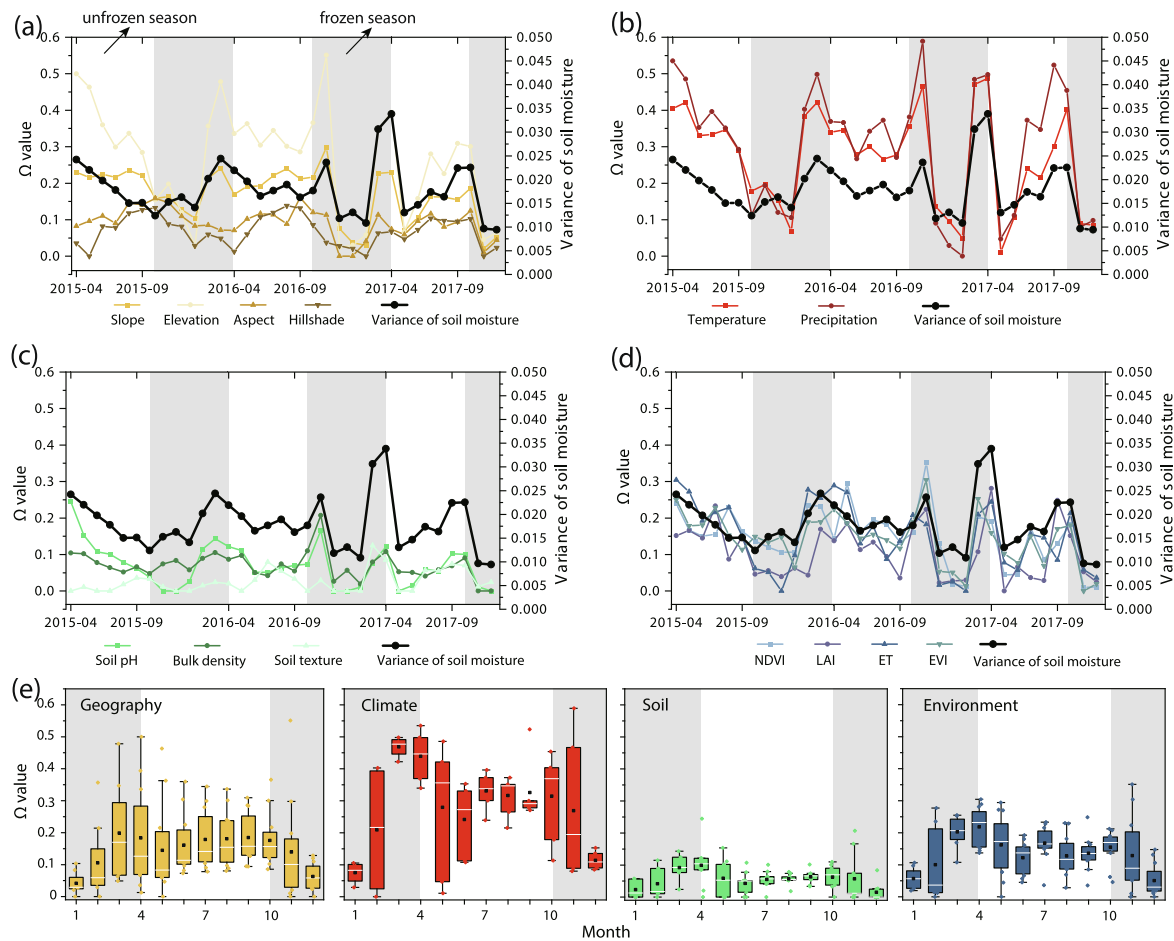


Fig. 4. Optimal power of determinants (OPD or Ω value) of explanatory variables on spatial patterns of monthly soil moisture. a: Ω values of geographical variables; b: Ω values of climate variables; c: Ω values of soil property variables; d: Ω values of environmental variables; and e: monthly summaries of Ω values from 2015 to 2017. White background shows unfrozen seasons (April–September), and gray background shows frozen seasons (October–March).

group (precipitation ≥ 0.094) contained zones B, C, and D, and the third group ($0.082 \leq \text{precipitation} < 0.094$) contained zones E, F, G, H, and I. The characteristics of soil moisture and explanatory variables in the three groups of zones are explained as follows.

The first group, including zone A, was primarily distributed in the western contiguous United States, Alaska, western Spain, southern France, and eastern Romania. The average precipitation in this group is only 0.047 mm/hr, which was much lower than the average precipitation in other groups, such as zone B (0.155 mm/hr) and zone F (0.087 mm/hr). The contiguous United States, southern France, and Romania were typical regions that zones were primarily divided by precipitation. In the contiguous United States, the western part was drought or desert areas and the precipitation was low, and the precipitation was gradually increased from the middle to eastern areas. In the southern France, the precipitation was low in the Mediterranean coast areas, but it was relatively high in the Massif Central areas (Planchon, 2000). The east of Romania was drought and most stations were distributed in zone A, but the western Romania was more humid than other areas and most of the stations were located in zone B.

The second group, including zones B, C, and D, was generally distributed in the eastern contiguous United States, Alaska, southern France, Denmark, western Germany, western Romania, northern Finland, and eastern Tibetan Plateau, China. Zones B, C, and D were divided by temperature, where the temperature in zone B was high, in zone C was low, and in zone D was moderate. For instance, soil moisture monitoring stations in western Germany were located in zones B and D. A typical characteristic of zone B was the high precipitation (> 0.094

mm/hr) and high temperature ($> 20^\circ\text{C}$), but the average temperature in zone D is 19.1°C , which was 30% lower than zone B (27.3°C).

The third group, including zones E, F, G, H, and I, was mainly located at Alaska, southwestern France, eastern Poland, and northeastern Finland. A few monitoring stations in this group were sparsely located at the contiguous United States, Hawaii, and southern Romania. We can find that zones E and F are usually distributed in neighbouring locations, such as central United States, southwestern France and southern Romania. The variable for differentiating zones E and F was NDVI, where the average NDVI in zone E and zone F was 0.49 and 0.65, respectively. For instance, in the southern Romania, the precipitation was moderate compared with eastern and western areas, and the stations were further divided into zones E and F by NDVI, where NDVI in zone F was 28% higher than that in zone E. Zones H and I had much higher average soil moisture than other zones, and they were divided by temperature with a threshold 22. For instance, stations in northeastern Poland were divided into zone H and I, and their average soil moisture were 0.58 and 0.42, respectively.

5.3.2. Frozen seasons

Fig. 7 shows the geographically optimal zones of soil moisture during the frozen season. The soil moisture monitoring stations were grouped into eleven spatial zones based on five variables using the GOZH model. The slope was the most important variable for determining the optimal zones. According to the slope value higher or lower than 0.33, soil moisture stations can be divided into two parts. Area with a slope value higher than 0.33 were generally mountainous areas. Compared with the

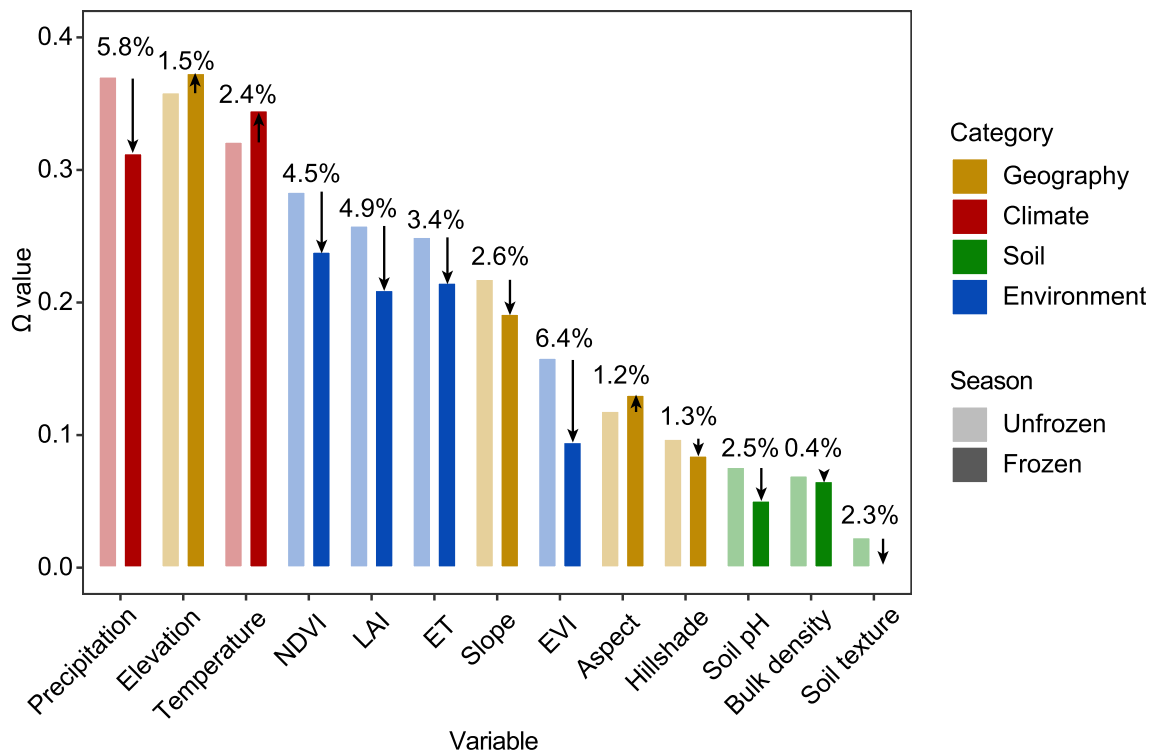


Fig. 5. Ω values of explanatory variables on spatial patterns of seasonal soil moisture and the comparison between unfrozen and frozen seasons.

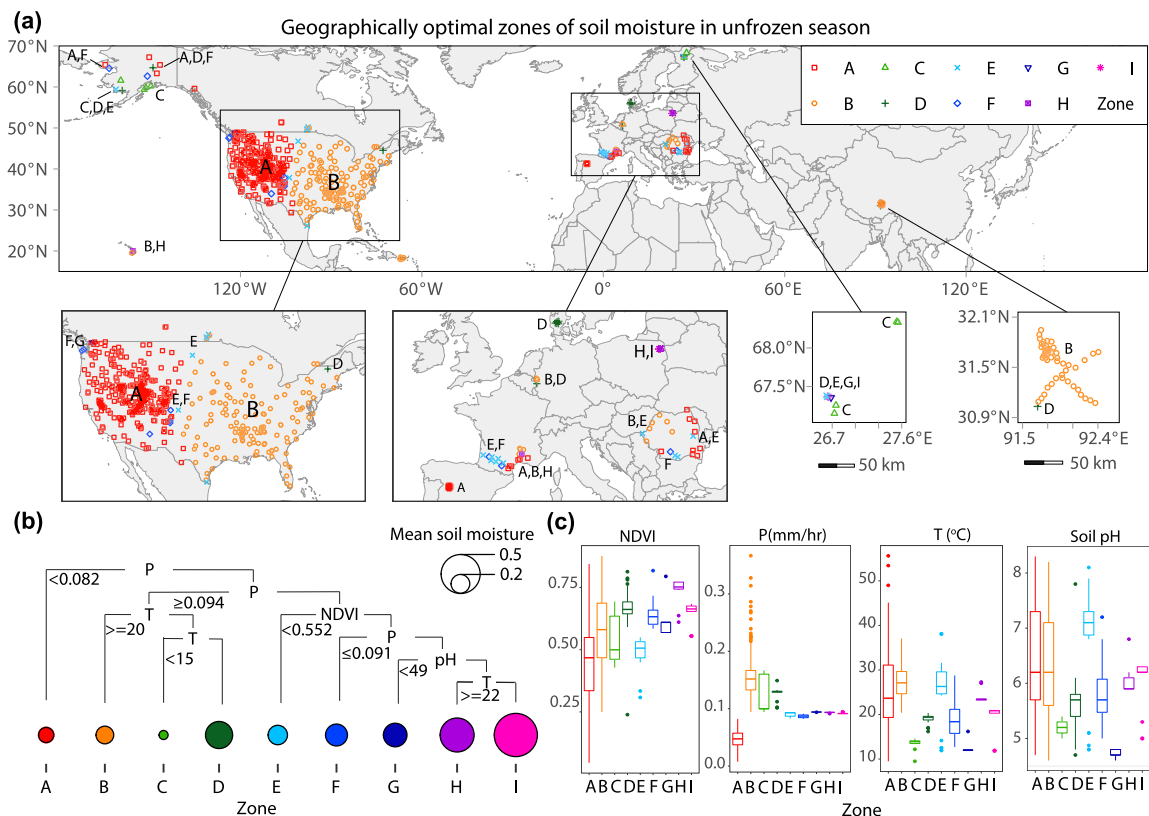


Fig. 6. Geographically optimal zones of soil moisture in the unfrozen season identified using the GOZH model (a), the process of identifying optimal zones (b), and statistical summaries of explanatory variables within zones for explaining characteristics of zones (c).

unfrozen season, terrain conditions had higher impacts on soil moisture disparities during the frozen season. In addition, ET and soil bulk density were variables that further divided zones in the second layers (Fig. 7 b).

According to variables in top two layers, including slope, ET, and soil bulk density, the eleven zones could be classified into four groups.

The first group, consisting of zones A and B, was characterized in

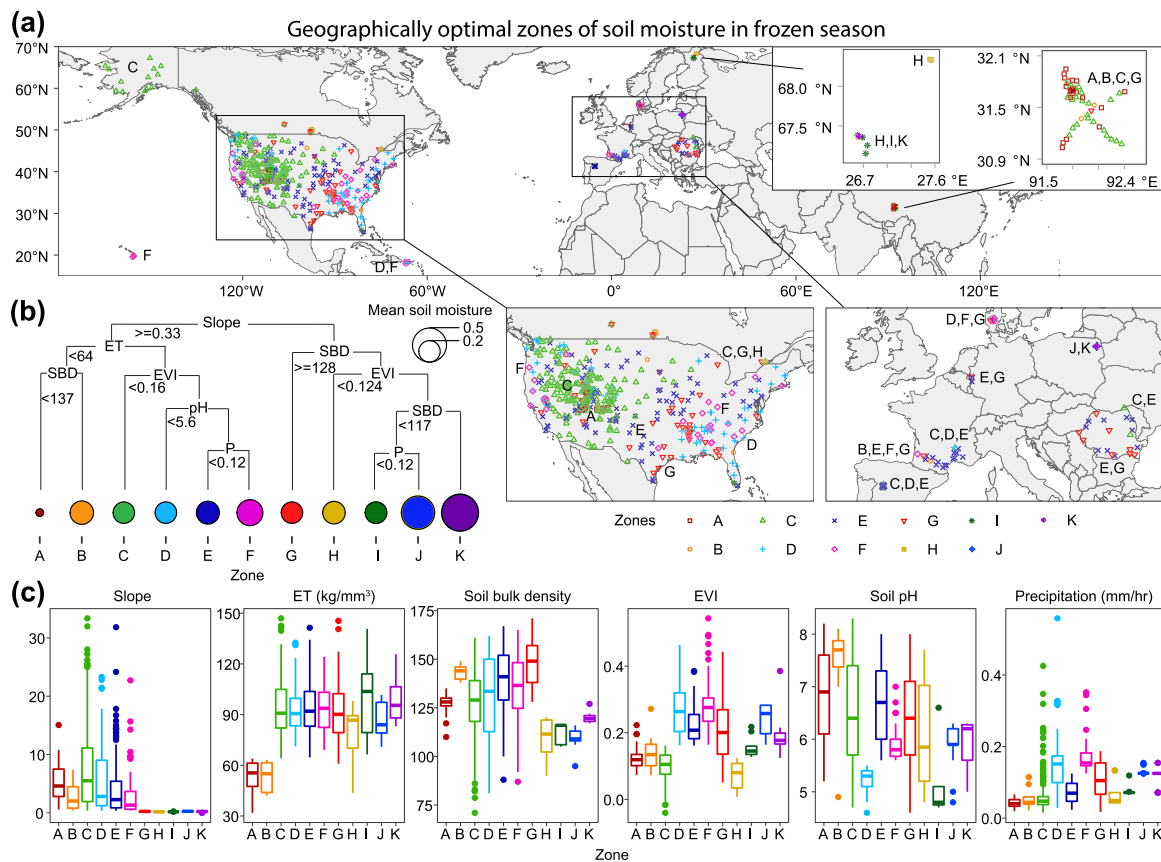


Fig. 7. Geographically optimal zones of soil moisture in the frozen season identified using the GPZH model (A), the process of identifying optimal zones (B), and statistical summaries of explanatory variables within zones for explaining characteristics of zones (C).

high slope and low ET. Zones A and B were divided by soil bulk density, where zone A had low soil bulk density and the lowest average soil moisture among all zones during the frozen season. Most of the stations in this group were located in the Rocky Mountains area in the United States and eastern Tibetan Plateau, China.

The second group, including zones C, D, E, and F, were also located in mountainous areas (slop > 0.33), but had relatively high ET ($ET > 64 \text{ kg/mm}^3$). For instance, in the mountainous areas of the western United States, most stations were located in zones C and E, where the average slope were 7.38 and 4.27, respectively. EVI was the variable dividing zones C and E. The average EVI in zone C and E was 0.098 and 0.30, respectively. This means that in addition to slope and ET, vegetation was an essential variable controlling the spatial variability of soil moisture in this region. In addition to the mountainous areas of the western United States, zone C was also distributed in Alaska, eastern Romania, and eastern Tibetan Plateau, China, and zone E was also located in the western Spain, southern France, western Germany, and Romania. In the western and eastern coastal areas of the United States, stations were located in zones D, E and zone F. They were divided by soil pH and precipitation. In zone D, soil pH was lower than 5.6 and the average pH was 0.53, but it was higher than 5.6 in zones E and F. Zones E and F were divided by the 0.12 mm/hr of the precipitation. The average precipitation in zone E and F was 0.107 and 0.17, respectively.

The third group, containing zone G, had low slope and high soil bulk density. Stations in this group were primarily located in the southern United States and Romania. The average slope in zone G is 0.20, which was much lower than its neighbouring zones, such as zone E (slope = 4.28).

The last group, including zones H, I, J, and K, had low slope and soil bulk density. Stations in this group were generally distributed in the northeastern Poland and northern Finland. For instance, stations in

northeastern Poland were divided into zones J and K. The soil bulk density controlled the spatial disparities in these two zones. The average soil bulk density in zones J and K were 111.55 and 109.44, respectively. In stations in the northern Finland, EVI and soil bulk density controlled spatial patterns of soil moisture.

5.4. Determinants of spatial disparities and seasonal effects

Table 3 shows overall Ω values of explanatory variables on spatial patterns of soil moisture investigated using the GOZH model and contributions of variables to the overall Ω values during unfrozen and frozen seasons. In general, overall Ω values were 47.62% and 47.69% during unfrozen and frozen seasons, respectively. This means that variables tended to have similar total contributions to spatial patterns of soil moisture during both seasons.

During the unfrozen season, climate variables had higher contributions to the overall Ω value, where contributions of precipitation and temperature were 20.99% and 7.90%, respectively. The contribution of precipitation accounted for 44.08% to the overall Ω value. In addition, NDVI and soil pH contributed 11.26% and 7.47%, respectively. All these contributions were lower than impacts of individual variables on spatial patterns of soil moisture. This means that explanatory variables had high interactive impacts on affecting patterns of soil moisture.

During the frozen season, spatial patterns of soil moisture were affected by slope, soil bulk density, ET, EVI, precipitation, and soil pH. The slope was closely associated with local terrain conditions, and it contributed 13.72% to patterns of soil moisture. Similar with the assessment of individual variables, geographical variables controlled the spatial variability of soil moisture during the frozen season. In addition to slope, soil bulk density, ET, EVI, precipitation, and soil pH contributed 12.62%, 6.58%, 6.31%, 4.04%, and 4.01% to spatial patterns of soil

Table 3

Contributions of explanatory variables on spatial patterns of soil moisture and contributions to dividing optimal zones in unfrozen and frozen seasons.

Unfrozen season			Frozen season		
Variable	Contribution to spatial patterns	Contribution to dividing zones	Variable	Contribution to spatial patterns	Contribution to dividing zones
Precipitation	20.99%	72.22%	Slope	13.72%	54.55%
NDVI	11.26%	11.11%	Soil bulk density	12.62%	12.73%
Temperature	7.90%	11.11%	ET	6.58%	14.55%
Soil pH	7.47%	5.56%	EVI	6.31%	10.91%
Overall Ω	47.62%	/	Precipitation	4.04%	3.64%
			Soil pH	4.01%	3.64%
			Overall Ω	47.69%	/

moisture during the frozen season, respectively.

In addition, Table 3 and Fig. 8 demonstrate contributions of variables to dividing optimal zones in unfrozen and frozen seasons. They had similar trends with contributions of variables to overall Ω values. For instance, precipitation contributed 72.22% to dividing zones during the unfrozen season, and slope contributed 54.55% to the division of zones during the frozen season.

5.5. Model evaluation

The performance of the GOZH model in investigating spatial heterogeneity in the large-scale soil moisture was evaluated from four aspects: exploring individual variables, assessing multiple spatial variables with interactive effects, dealing with finely divided zones during spatial overlay, and the reliability of models. These aspects of the GOZH model were evaluated by comparing with the commonly used OPGD model during the unfrozen and frozen seasons. Fig. 9 a-d shows the spatial discretization process of the OPGD model for 12 variables, in addition to soil texture, which was a categorical variable containing six classes of texture. With the break number increase from 1 to 16, the PD, i.e., Q value, of all variables increased gradually. The optimal break numbers were selected when the increase rate was lower than 0.05. In this study, 9 and 12 were selected as the optimal break numbers of continuous variables during unfrozen and frozen seasons, respectively. Fig. 9 e and f shows OPGD-based PD values of individual variables to spatial patterns of soil moisture.

First, the GOZH model supports the derivation of the maximum spatial associations between response and explanatory variables through the identification of geographically optimal zones. The maximum spatial associations can accurately reveal the spatial

heterogeneity of soil moisture. Therefore, the GOZH model is a reliable approach for examining spatial heterogeneity and exploring OPD of explanatory variables on spatial patterns of soil moisture.

In addition, the GOZH model can help reduce the underestimation of the PD values by the OPGD model as demonstrated by explorations of individual variables. Ranks of PD values explored by OPGD models during both unfrozen and frozen seasons were similar to those of GOZH models. For instance, precipitation and elevation were variables with both the highest Q (PD) and Ω (OPD) values during unfrozen and frozen seasons, respectively. However, the power of explanatory variables revealed by the GOZH model had a significant enhancement than the OPGD model. The average Ω values of individual variables were 80.9% and 68.2% higher than the average Q values during the unfrozen and frozen seasons, respectively.

Third, the GOZH model can effectively avoid the overestimation of the interactive impacts of multiple spatial variables on patterns of soil moisture compared with the OPGD model. Fig. 10 shows a model performance comparison between the GOZH and OPGD models in terms of the OPD/PD of variables and numbers of zones with the increased number of explanatory variables during the unfrozen and frozen seasons. In Fig. 10 a and c, average GOZH-based Ω values of individual variables were 0.20 and 0.18, and they were gradually increased to 0.48. In GOZH models, numbers of zones were not critically increased (Fig. 10 b and d), which indicated the robustness of GOZH models in the analysis of spatial heterogeneity. However, in Fig. 10 e and g, average OPGD-based Q values of individual variables were both 0.11, respectively, and they were rapidly increased to 0.99. Simultaneously, numbers of zones were also critically increased from 13 to 750 when the number of variables was higher than 1. The critically increased number of zones caused the very limited observations within zones and made the

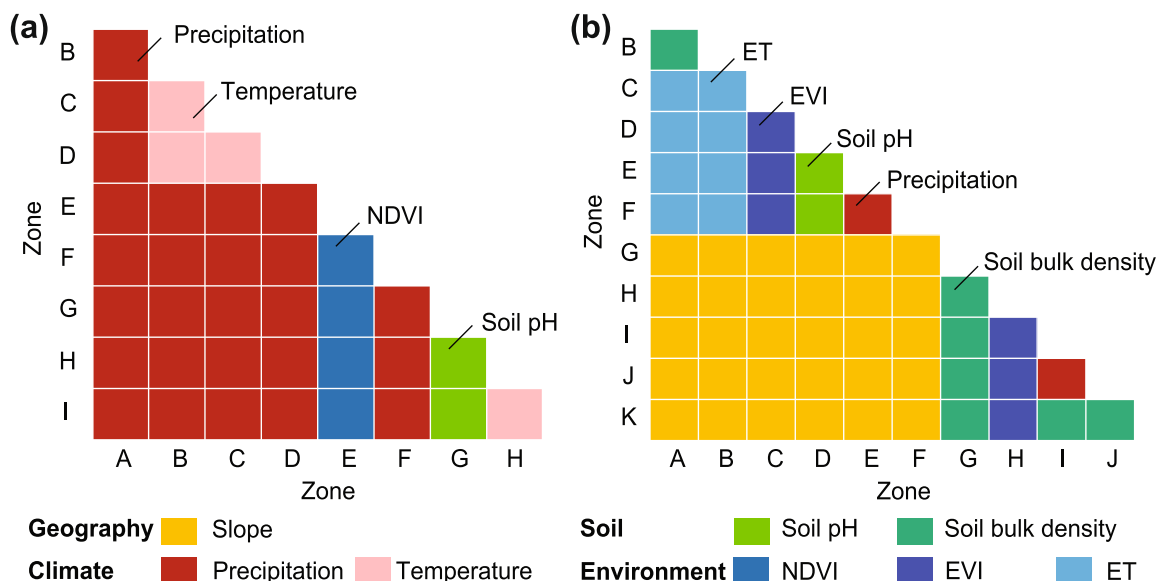


Fig. 8. Summary of explanatory variables used for dividing each pair of geographical optimal zones of soil moisture in unfrozen (a) and frozen (b) seasons.

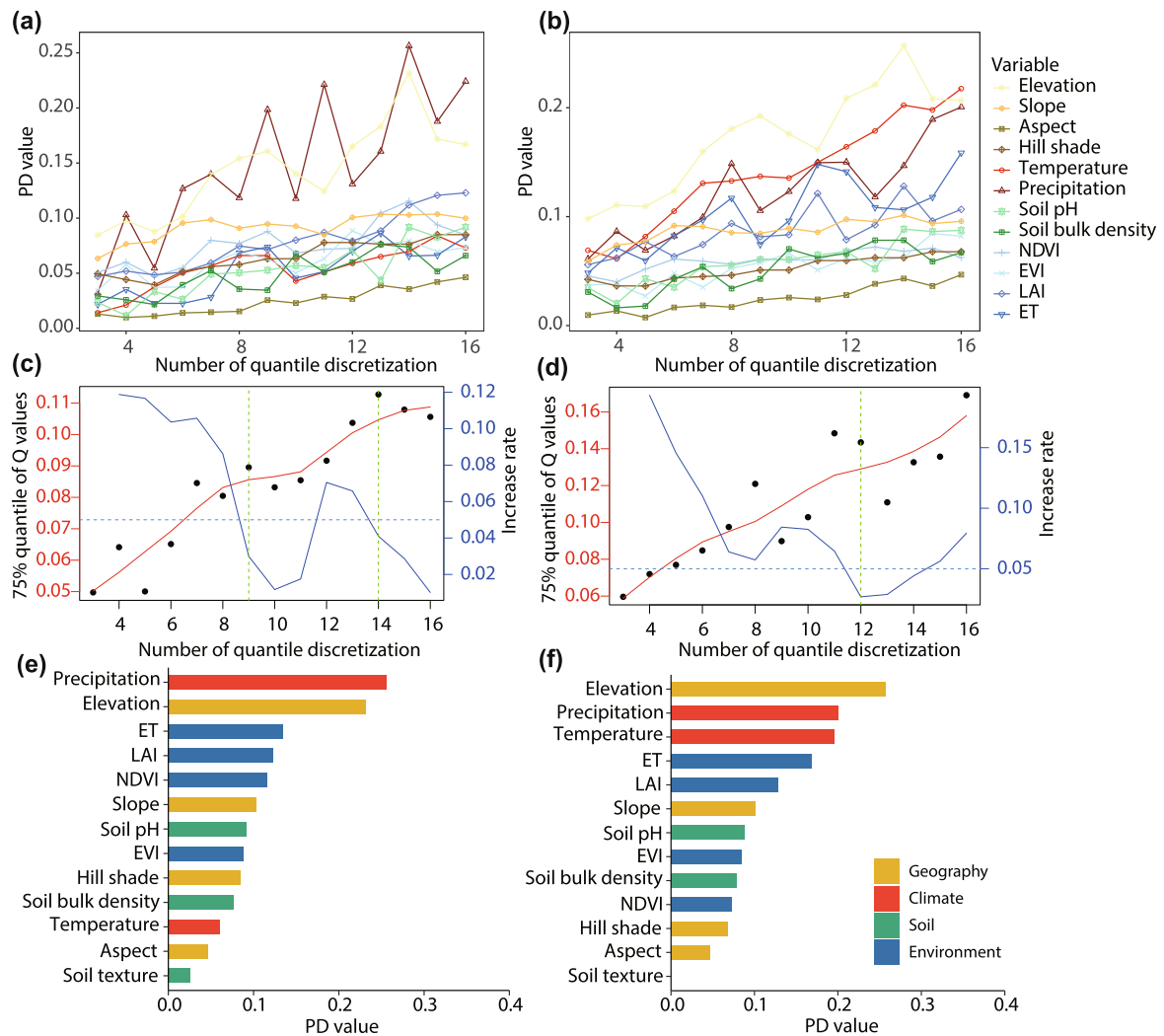


Fig. 9. Processes and results of the optimal parameters-based geographical detectors (OPGD) model for assessing power of determinants (PD) of soil moisture. PD of variables with different numbers of spatial discretization (a and b), processes of selecting optimal numbers of discretization (c and d), and PD of individual variables using optimal parameters (e and f) in unfrozen and frozen seasons, respectively.

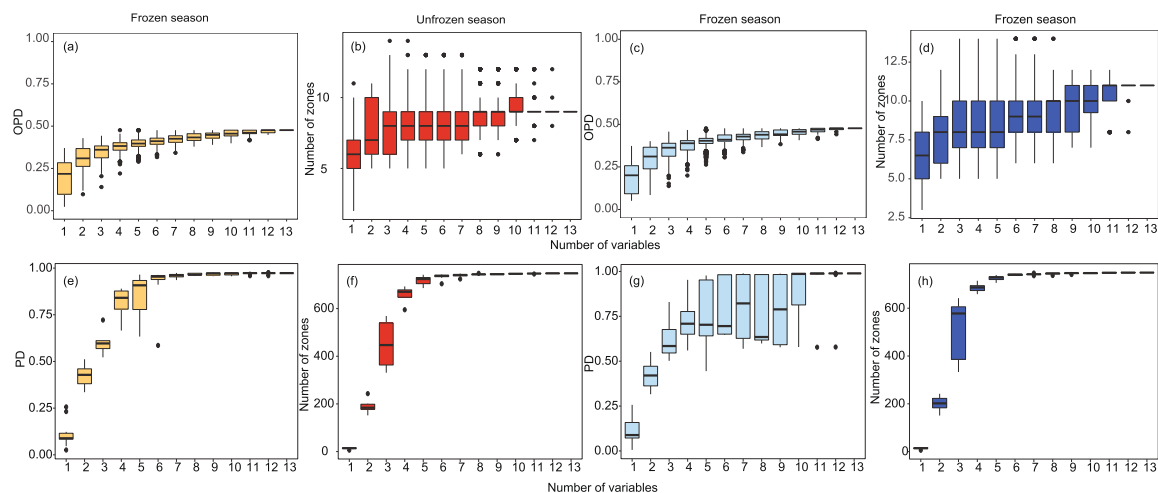


Fig. 10. Model performance comparison between GOZH and OPGD models: power of determinants (PD) and number of zones in the unfrozen and frozen seasons investigated using the GOZH model (a-d), and that investigated using the OPGD model (e-h).

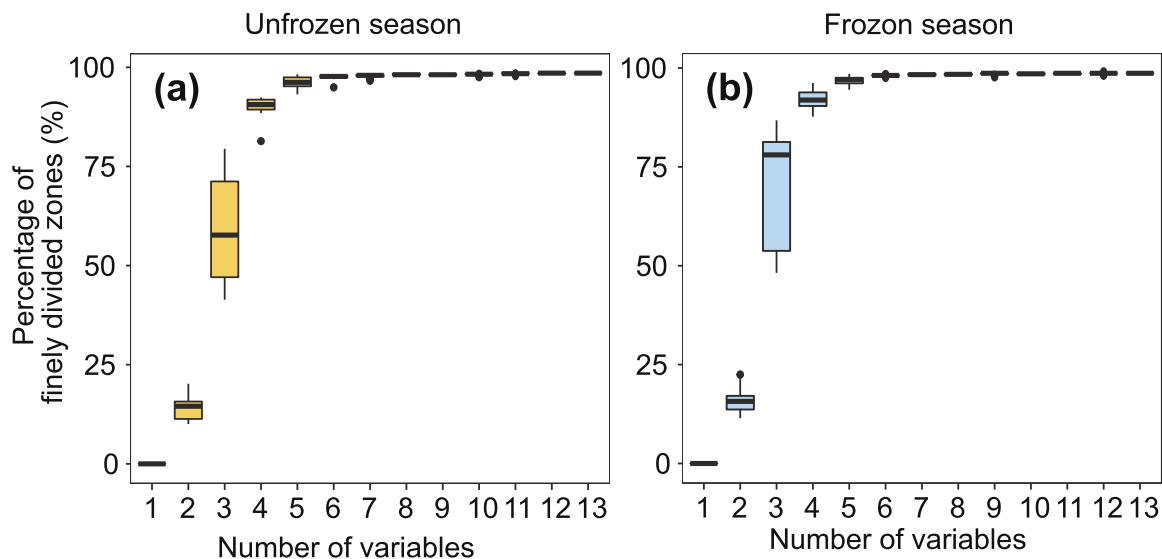


Fig. 11. Percentages of finely divided zones (FDZ), which are zones contain only an observation, are critically increased in the OPGD models for unfrozen (A) and frozen (B) seasons, when the number of variables is higher than two. On the contrary, percentages of FDZ in GOZH-based results are zero for both seasons.

increased Q values unreliable.

Finally, the GOZH model can eliminate finely divided zones (FDZs) that are common in the interactive variable assessment of SSH models. Fig. 11 shows the percentages of FDZs in OPGD models. In OPGD models, when examining interactive impacts of any two variables on soil moisture patterns, about 20% of zones would be FDZs, where there was only one observation in each FDZs. When the number of variables was higher than 5, nearly all zones would be FDZs. The extremely high percentage of FDZs cannot explain the real PD of variables. Thus, the analysis of FDZs demonstrates that the rapidly increased Q values of the interaction of multiple variables in OPGD models were not reliable when three or more variables were used in models. The analysis also explains why only two variables were considered in the interaction analysis in previous OPGD-based studies. When the variable number exceeds 2, most of the zones only have one or a few observations, which makes it difficult to reveal the real PD of the interaction of variables. However, in GOZH models, no FDZs existed no matter how many explanatory variables were used. The analysis of FDZs can further confirm the reliability and robustness of the GOZH model in the analysis of spatial heterogeneity.

6. Discussion

6.1. Methodological contributions

This study proposed a GOZH model to explore the spatial variability of soil moisture in the Northern Hemisphere. The GOZH model has following advantages in spatial determinant explorations. First, in the GOZH model, an OPD indicator was developed reveal the maximum spatial associations between soil moisture variability and determinants. Second, the optimal geographical zones can be derived from the explanatory variables. Third, no statistical assumptions are required in the GOZH model. Finally, the model validation in the study has demonstrated that the GOZH model can effectively explore spatial determinants of soil moisture through avoiding the underestimation of individual variables, overestimation of multiple variables, and finely divide zones.

6.2. Complex spatial heterogeneity of soil moisture patterns

This study revealed the complex spatial heterogeneity of soil moisture during unfrozen and frozen season at a global scale. The complexity

of spatial heterogeneity of soil moisture patterns can be explained in following aspects. First, spatial patterns of soil moisture had significant regional disparities that were closely associated with regional geographical, climate, soil, and environmental conditions. The soil moisture monitoring stations in the Northern Hemisphere can be divided into nine and eleven zones during the unfrozen and frozen seasons, respectively. Explanatory variables tended to be similar within zones, and significantly varied among different zones.

In addition, determinants of spatial patterns of large-scale soil moisture have seasonal characteristics. On one hand, the spatial heterogeneity during unfrozen and frozen seasons has similarities. The overall Ω values that examined the maximum PD of four categories of explanatory variables in both seasons were approximate 48%, and they were affected by the interaction of multiple spatial variables. These results revealed the complexity of spatial heterogeneity in soil moisture, that only half of the heterogeneity could be explained by geographical, climate, soil, and environmental variables.

On the other hand, the spatial heterogeneity of soil moisture during the frozen season was more complex than that during the unfrozen season. First, spatial distributions of geographically optimal zones during the frozen season was much more complex than those during the unfrozen season, which appeared in most monitoring stations in the North America, Europe, and China. An exception was Alaska, where stations were divided into five zones during the unfrozen season, but they were located in a zone during the frozen season. Second, more explanatory variables are required to identify geographical optimal zones and estimate Ω values, where numbers of required explanatory variables during the unfrozen and frozen seasons were four and six, respectively. Finally, climate variables, including precipitation and temperature, are predominant variables of spatial patterns of soil moisture during the unfrozen season, accounting for 60.7% of the overall Ω value, but spatial patterns of soil moisture were affected by all four categories of variables during the unfrozen season, where the primary variable, slope, only accounted for 28.8% of the overall Ω value. Precipitation contributed 72.2% to dividing zones. During the frozen season, geographical variables controlled the spatial variability of soil moisture, and slope contributed 54.55% to dividing zones. Soil properties, including soil bulk density and soil pH also determines the geographical optimal zones during the frozen season. The slope can explain 19.16% of soil moisture during the frozen season, which is lower than its contribution to the geographically optimal zones combining with other variables. This result is consistent with previous works, which

geographical variables can not be the single determinant of soil moisture variability (Wilson et al., 2005).

Finally, it is more difficult to characterize and explain the spatial heterogeneity than temporal heterogeneity in soil moisture. For instance, previous studies have demonstrated that the fitness of temporal prediction of soil moisture could reach to 96% using deep learning models (Cai et al., 2019; Ahmed et al., 2021). However, the accuracy of spatial prediction was much lower than that in temporal predictions and the accuracy tended to decrease with the increased spatial scale. For instance, the fitness of spatial prediction at local scales were 0.41–0.84 in multiple studies (Badewa et al., 2018; Peng et al., 2017), and that at a global scale can only reached to 0.63 (Montzka et al., 2018), even models have been improved with the consideration of more soil information and characteristics, such as soil texture (Montzka et al., 2018). According to this study, the primary reason of the lower accuracy of spatial prediction is the complex spatial heterogeneity that only about a half of the spatial heterogeneity of soil moisture can be explained by explanatory variables.

6.3. Temporal variations of soil moisture determinants

The temporal phase of soil moisture determinants was investigated from following three aspects. First, the spatial heterogeneity and determinants was varied from April 2015 to December 2017. Generally, climate variables, i.e. precipitation and temperature, have the highest spatial associations with soil moisture during most periods but the associations were varied in different months, which is consistent with previous findings (Wang et al., 2017). Explanatory variables had the highest explanatory power to soil moisture in November 2016, where precipitation can impact 58% of soil moisture, and elevation and temperature can explain 55% and 46% of the spatial variability of soil moisture. However, variables can only explain up to 5% of soil moisture in February 2017. Therefore, a relatively long time period, such as a half year, is recommended for reliable explorations of variability and determinants of spatial patterns of soil moisture.

Second, soil moisture spatial variability has a strong monthly pattern. In the frozen-unfrozen season-changing months, i.e., March and April, spatial associations between patterns of soil moisture and explanatory variables generally have the highest Ω value, and the Ω value of climate variables have the highest improvement in this period. Previous studies have concluded that during the transition phase, climate variables become more important to soil moisture variability likely because the alternation of cold and warm days controlled by weather variability (Kang et al., 2010; Wei et al., 2019). The least interpretable period of spatial patterns of soil moisture is the middle frozen seasons.

Third, seasonal effects were identified in the spatial heterogeneity of soil moisture, results show that the spatial pattern of soil moisture is more interpretable during the unfrozen season than during the frozen season. The average Ω value of individual variables during the unfrozen season (20.0%) is 12.4% higher than that during the frozen season (17.8%). Different from most variables, Ω values of temperature and two geographical variables, elevation and aspect, are increased from unfrozen to frozen season. This finding is consistent with previous research that in cold weather, soil moisture variability is strongly associated with global warming, and the impacts of temperature can be more significant (Kang et al., 2010; Wei et al., 2019). Studies also found geographical variables is the main driver of soil moisture during the winter when soil is frozen which particular because its association to water table (Rosebaum et al., 2012). During the unfrozen season, the impacts of geographical variables are negligible due to the low water table (Chaney et al., 2015). During the frozen season, with the low ET, the water table increases and closes to the surface, which enables higher impacts of the groundwater and subsurface flow on the soil moisture variability (Western et al., 1998; Rosenbaum et al., 2012).

6.4. Contributions of heterogeneity and geographical zones to soil moisture studies

Findings about the spatial heterogeneity of soil moisture in this study can help optimize the design of soil moisture monitoring network, spatial down-scaling of soil moisture data, and accurate inversion of surface parameters from soil moisture.

First, the network design of soil moisture can be optimized with the improved understanding of the spatial heterogeneity and determinants of regional disparities of soil moisture identified in this study. Due to the complex heterogeneity of spatial soil moisture, most existing in situ observation networks rarely provide sufficient coverage to capture soil moisture variability at a watershed scale. Thus, it is critically required to develop a systematic approach to soil moisture network design in order to accurately capture soil moisture information in the watershed space with a minimum number of sensors. It was found that the current (simulated and observed) network of soil moisture detectors underestimates the average spatial heterogeneity (Zhuo et al., 2020). The analysis of the determinants of soil moisture heterogeneity and the spatial partitioning results from the GOZH model can be used to inform the development of new techniques for ground-based measurement network design. The intended network design can take into account the spatial variability of soil moisture.

Second, the spatial down-scaling of the soil moisture data requires the spatial heterogeneity information of large-scale soil moisture monitoring data. The coarse resolution of soil moisture remote sensing products limits its application at fine scales, which introduces the need for their spatial down-scaling (Chaney et al., 2015). A series of down-scaling methods had been developed to improve resolutions of soil moisture products using multi-source auxiliary data and various methods, such as statistical models, geospatial models, machine learning, deep learning, and hybrid models (Peng et al., 2017). However, due to the existence of spatial heterogeneity of soil moisture, the accuracy of spatial prediction has been lower than that of temporal prediction (Badewa et al., 2018; Peng et al., 2017; Montzka et al., 2018). It is also a challenge to quantitatively assess the large differences in soil moisture determinants in different regions (Molero et al., 2018). The geographically optimal zones of soil moisture obtained using the GOZH model, and the control factors in different regions can effectively guide the spatial down-scaling process. Our study shows that geographical variables are the most important factors to soil moisture in the frozen season. For example, soil moisture heterogeneity in the east and west of North America is controlled by the slope. Therefore, greater weight should be given to geographical variables during spatial down-scaling. In the unfrozen season, environmental and climate variables are essential to soil moisture. Precipitation determines the soil moisture in the western United States, while NDVI determines soil moisture in the central United States.

Finally, understanding soil moisture heterogeneity over different geographical zones can also support the accurate inversion of surface parameters from soil moisture satellite data. The limited knowledge of regional differences in soil moisture and its determinants poses a challenge to calibrate ground roughness parameterization schemes with ground observation data. Obtaining information on soil moisture heterogeneity can improve the accuracy and the geographical transferability of the parameterization scheme. (Verhoest et al., 2008).

There are still limitations of this study. First, the scale effect between soil moisture in situ data and remote sensing images were not considered in this study. The explanatory variables are derived from the pixels in the position of the soil moisture monitor stations. Spatial heterogeneity of soil moisture at stations in the Northern Hemisphere is much higher than that of data within grids of explanatory variables, e.g., 90 m or 250 m. Therefore, we assume spatial analysis in the study will not be affected by the scale effects of explanatory variables derived from remote sensing or grid data. In addition, some explanatory variables, for example, elevation and slope, may be represented by a zone with an area larger

than the size of grid in the images (Jasiewicz and Stepinski, 2013). In this case, data at surrounding grids need to be considered for deriving explanatory variables at stations. From the perspective of spatial heterogeneity models, approaches can be developed for more effective use of continuous variables in spatial heterogeneity models. For instance, the spatial association detector (SPADE) (Cang and Luo, 2018) and the interactive detector for spatial association (IDSA) (Song and Wu, 2021) models were developed to compare zonal and global spatial dependence, i.e., spatial autocorrelation of data, instead of zonal and global variance, for computing the PD values. The K-means (Likas et al., 2003; Hartigan and Wong, 1979) and hierarchical clustering (Johnson, 1967) methods also can be used to derive spatial zones with the continuous explanatory variables. Finally, the division of the frozen and unfrozen seasons in this study may introduce uncertainty. The study aims to explore the soil moisture variability in the Northern Hemisphere. The frozen/ unfrozen months were unified in the whole study area since most stations are located in the mid-latitude area and only thawed soil moisture data were selected and analyzed. However, some stations are located in the high latitude area like Alaska, where the unfrozen/frozen season of soil moisture may be different from other areas. Thus, further studies may explore the soil moisture variability in different climate zones.

7. Conclusion

This study developed a geographically optimal zones-based heterogeneity (GOZH) model to explore the spatial variability of soil moisture in the Northern Hemisphere. In the GOZH model, the optimal power of determinant (OPD) indicator can reveal the maximum spatial associations, and the spatial determinants can be effectively explored through avoiding the underestimation of individual variables, overestimation of multiple variables, and finely divide zones.

The GOZH model was implemented to explore the spatial and temporal patterns of soil moisture variability. Results shows that in the frozen-unfrozen season-changing months, spatial associations between patterns of soil moisture and explanatory variables generally have the highest OPD value especially for climate variables. The average OPD value of individual variables during the unfrozen season (20.0%) is higher than that during the frozen season (17.8%). In addition, geographically optimal zones and corresponding determinants of soil moisture were revealed by the interactive of explanatory variables. Variables have similar contributions to spatial pattern of soil moisture during two seasons. At a global scale, the combinations of determinants can explain about 48% of the spatial pattern of soil moisture. During the unfrozen season, climate variables, including precipitation and temperature, have the highest contributions to the overall OPD value. During the frozen season, geographical variables (e.g., slope) controlled the spatial variability of soil moisture.

This study can provide a deep understanding of variability and determinants of soil moisture at a global scale. The knowledge of soil moisture determinants can be better used in situ network design, spatial down-scaling of soil moisture. In addition, the results can also be applied to the evaluate soil moisture in satellite imagery and the accurate inversion of surface parameters from satellite data on soil moisture.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by funding from the Curtin University and the China Scholarship Council.

References

- Ahmed, A., Deo, R.C., Raj, N., Ghahramani, A., Feng, Q., Yin, Z., Yang, L., 2021. Deep learning forecasts of soil moisture: Convolutional neural network and gated recurrent unit models coupled with satellite-derived modis, observations and synoptic-scale climate index data. *Remote Sens.* 13 (4), 554.
- Al-Yaari, A., Wigneron, J.P., Dorigo, W., Colliander, A., Pellarin, T., Hahn, S., Mialon, A., Richaume, P., Fernandez-Moran, R., Fan, L., Kerr, Y.H., De Lannoy, G., 2019. Assessment and inter-comparison of recently developed/reprocessed microwave satellite soil moisture products using ISMN ground-based measurements. *Remote Sens. Environ.* 224 (February), 289–303.
- Albergel, C., Calvet, J.-C., Rosnay, P. d., Balsamo, G., Wagner, W., Hasenauer, S., Naeimi, V., Martin, E., Bazile, E., Bouyssel, F., et al., 2010. Cross-evaluation of modelled and remotely sensed surface soil moisture with in situ data in southwestern France. *Hydrol. Earth Syst. Sci.* 14 (11), 2177–2191.
- Albergel, C., De Rosnay, P., Gruhier, C., Muñoz-Sabater, J., Hasenauer, S., Isaksen, L., Kerr, Y., Wagner, W., 2012. Evaluation of remotely sensed and modelled soil moisture products using global ground-based in situ observations. *Remote Sens. Environ.* 118, 215–226.
- Albergel, C., Rüdiger, C., Carrer, D., Calvet, J.-C., Fritz, N., Naeimi, V., Bartalis, Z., Hasenauer, S., 2009. An evaluation of ASCAT surface soil moisture products with in-situ observations in Southwestern France. *Hydrol. Earth Syst. Sci.* 13 (2), 115–124 <https://hess.copernicus.org/articles/13/115/2009/>.
- Albergel, C., Rüdiger, C., Pellarin, T., Calvet, J.-C., Fritz, N., Froissard, F., Suquia, D., Petitpa, A., Pignet, B., Martin, E., 2008. From near-surface to root-zone soil moisture using an exponential filter: an assessment of the method based on in-situ observations and model simulations. *Hydrol. Earth Syst. Sci.* 12 (6), 1323–1337.
- Babaeian, E., Sadeghi, M., Franz, T.E., Jones, S., Tuller, M., 2018. Mapping soil moisture with the optical trapezoid model (optram) based on long-term modis observations. *Remote Sens. Environ.* 211, 425–440.
- Badewa, E., Unc, A., Cheema, M., Kavanagh, V., Galagedara, L., 2018. Soil moisture mapping using multi-frequency and multi-coil electromagnetic induction sensors on managed podzols. *Agronomy* 8 (10), 224.
- Baroni, G., Ortuani, B., Facchi, A., Gandolfi, C., 2013. The role of vegetation and soil properties on the spatio-temporal variability of the surface soil moisture in a maize-cropped field. *J. Hydrol.* 489, 148–159.
- Bell, J.E., Palecki, M.A., Baker, C.B., Collins, W.G., Lawrimore, J.H., Leeper, R.D., Hall, M.E., Kochendorfer, J., Meyers, T.P., Wilson, T., et al., 2013. Us climate reference network soil moisture and temperature observations. *Journal of Hydrometeorology* 14 (3), 977–988.
- Bell, K.R., Blanchard, B., Schumge, T., Witzczak, M., 1980. Analysis of surface moisture variations within large-field sites. *Water Resour. Res.* 16 (4), 796–810.
- Berg, A., Sheffield, J., Milly, P.C., 2017. Divergent surface and total soil moisture projections under global warming. *Geophys. Res. Lett.* 44 (1), 236–244.
- Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 2017. *Classification and regression trees*. Routledge.
- Brocca, L., Melone, F., Moramarco, T., Morbidelli, R., 2010. Spatial-temporal variability of soil moisture and its estimation across scales. *Water Resour. Res.* 46 (2).
- Brocca, L., Melone, F., Moramarco, T., Wagner, W., Hasenauer, S., 2010. Ascet soil wetness index validation through in situ and modeled soil moisture data in central Italy. *Remote Sens. Environ.* 114 (11), 2745–2755.
- Cai, Y., Zheng, W., Zhang, X., Zhangzhong, L., Xue, X., 2019. Research on soil moisture prediction model based on deep learning. *PLoS one* 14 (4), e0214508.
- Cang, X., Luo, W., 2018. Spatial association detector (SPADE). *International Journal of Geographical Information Science* 32 (10), 2055–2075. <https://doi.org/10.1080/13658816.2018.1476693>.
- Cang, X., Luo, W., 2018. Spatial association detector (spade). *International Journal of Geographical Information Science* 32 (10), 2055–2075.
- Chaney, N.W., Roundy, J.K., Herrera-Estrada, J.E., Wood, E.F., 2015. High-resolution modeling of the spatial heterogeneity of soil moisture: Applications in network design. *Water resources research* 51 (1), 619–638.
- Dari, J., Morbidelli, R., Saltalippi, C., Massari, C., Brocca, L., 2019. Spatial-temporal variability of soil moisture: Addressing the monitoring at the catchment scale. *Journal of Hydrology* 570 (October 2018), 436–444.
- Chen, R., Yan, D., Wen, A., Shi, Z., Chen, J., Liu, Y., Chen, T., 2021. The regional difference in engineering-control and tillage factors of Chinese Soil Loss Equation. *J. Mount. Sci.* 18 (3), 658–670.
- Das, N.N., Mohanty, B.P., 2008. Temporal dynamics of psr-based soil moisture across spatial scales in an agricultural landscape during smex02: A wavelet approach. *Remote Sens. Environ.* 112 (2), 522–534.
- Didan, K., Munoz, A.B., Solano, R., Huete, A., 2015. MODIS Vegetation Index User 's Guide (Collection 6) 2015 (May), 31.
- Dorigo, W., Himmelbauer, I., Aberer, D., Schremmer, L., Petrakovic, I., Zappa, L., Preimesberger, W., Xaver, A., Annor, F., Ardö, J., Baldocchi, D., Blöschl, G., Boga, H., Brocca, L., Calvet, J.-C., Camarero, J., Capello, G., Choi, M., Cosh, M., Demarty, J., van de Giesen, N., Hajdu, I., Jensen, K., Kanniah, K.D., de Kat, I., Kirchengast, G., Rai, P.K., Kyrouac, J., Larson, K., Liu, S., Loew, A., Moghaddam, M., Martínez Fernández, J., Mattar Bader, C., Morbidelli, R., Musial, J., Osenga, E., Palecki, M., Pfeil, I., Powers, J., Ikonen, J., Robock, A., Rüdiger, C., Rummel, U., Strobel, M., Su, Z., Sullivan, R., Tagesson, T., Vreugdenhil, M., Walker, J., Wigneron, J.P., Woods, M., Yang, K., Zhang, X., Zreda, M., Dietrich, S., Gruber, A., van Oevelen, P., Wagner, W., Scipal, K., Drusch, M., Sabia, R., 2021. The International Soil Moisture Network: serving Earth system science for over a decade. *Hydrology and Earth System Sciences Discussions* (January), 1–83.
- Dorigo, W.A., Gruber, A., De Jeu, R.A.M., Wagner, W., Stacke, T., Loew, A., Albergel, C., Brocca, L., Chung, D., Parinussa, R.M., Kidd, R., 2015. Evaluation of the ESA CCI soil

- moisture product using ground-based observations. *Remote Sens. Environ.* 162, 380–395 <https://www.sciencedirect.com/science/article/pii/S0034425714002727>.
- Dorigo, W.A., Wagner, W., Hohensinn, R., Hahn, S., Paulik, C., Xaver, A., Gruber, A., Drusch, M., Mecklenburg, S., van Oevelen, P., Robock, A., Jackson, T., 2011. The International Soil Moisture Network: a data hosting facility for global in situ soil moisture measurements. *Hydrol. Earth Syst. Sci.* 15 (5), 1675–1698 <https://hess.copernicus.org/articles/15/1675/2011/>.
- Elkhrachy, I., 2018. Vertical accuracy assessment for SRTM and ASTER Digital Elevation Models: A case study of Najran city, Saudi Arabia. *Ain Shams Engineering Journal* 9 (4), 1807–1817 <https://www.sciencedirect.com/science/article/pii/S2090447917300084>.
- Entekhabi, D., Njoku, E.G., O'Neill, P.E., Kellogg, K.H., Crow, W.T., Edelstein, W.N., Entin, J.K., Goodman, S.D., Jackson, T.J., Johnson, J., et al., 2010. The soil moisture active passive (smap) mission. *Proc. IEEE* 98 (5), 704–716.
- Entin, J.K., Robock, A., Vinnikov, K.Y., Hollinger, S.E., Liu, S., Namkhai, A., 2000. Meteorologic i Land Surface. *J. Geophys. Res.* 105 (D9), 11865–11877.
- Famiglietti, J., Wood, E.F., 1994. Multiscale modeling of spatially variable water and energy balance processes. *Water Resour. Res.* 30 (11), 3061–3078.
- Famiglietti, J.S., Ryu, D., Berg, A.A., Rodell, M., Jackson, T.J., 2008. Field observations of soil moisture variability across scales. *Water Resour. Res.* 44 (1).
- Fang, H., Baret, F., Plummer, S., Schaepman-Strub, G., 2019. An Overview of Global Leaf Area Index (LAI): Methods, Products, Validation, and Applications. *Rev. Geophys.* 57 (3), 739–799.
- Green, J.K., Seneviratne, S.I., Berg, A.M., Findell, K.L., Hagemann, S., Lawrence, D.M., Gentile, P., 2019. Large influence of soil moisture on long-term terrestrial carbon uptake. *Nature* 565 (7740), 476–479.
- Gruber, A., Dorigo, W.A., Zwieback, S., Xaver, A., Wagner, W., 2013. Characterizing Coarse-Scale Representativeness of in situ Soil Moisture Measurements from the International Soil Moisture Network. *Vadose Zone Journal* 12 (2). <https://doi.org/10.2136/vzj2012.0170>.
- Han, J., Mao, K., Xu, T., Guo, J., Zuo, Z., Gao, C., 2018. A soil moisture estimation framework based on the cart algorithm and its application in china. *Journal of hydrology* 563, 65–75.
- Hartigan, J.A., Wong, M.A., 1979. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)* 28 (1), 100–108.
- Hashimoto, H., Dungan, J.L., White, M.A., Yang, F., Michaelis, A.R., Running, S.W., Nemani, R.R., 2008. Satellite-based estimation of surface vapor pressure deficits using modis land surface temperature data. *Remote Sens. Environ.* 112 (1), 142–155.
- Holzman, M.E., Rivas, R., Bayala, M., 2014. Subsurface soil moisture estimation by vi-1st method. *IEEE Geosci. Remote Sens. Lett.* 11 (11), 1951–1955.
- Hou, A.Y., Kakar, R.K., Neeck, S., Azarbarzin, A.A., Kummerow, C.D., Kojima, M., Oki, R., Nakamura, K., Iguchi, T., 2014. The global precipitation measurement mission. *Bull. Am. Meteorol. Soc.* 95 (5), 701–722.
- Hu, W., Shen, Q., Zhai, X., Du, S., Zhang, X., 2021. Impact of environmental factors on the spatiotemporal variability of soil organic matter: a case study in a typical small Mollisol watershed of Northeast China. *J. Soils Sediments* 21 (2), 736–747.
- Jasiewicz, J., Stepinski, T.F., 2013. Geomorphons—a pattern recognition approach to classification and mapping of landforms. *Geomorphology* 182, 147–156.
- Jawson, S.D., Niemann, J.D., 2007. Spatial patterns from eof analysis of soil moisture at a large scale and their dependence on soil, land-use, and topographic properties. *Adv. Water Resour.* 30 (3), 366–381.
- Jensen, K.H., Illangasekare, T.H., 2011. Hobe: A hydrological observatory. *Vadose Zone Journal* 10 (1), 1–7.
- Jiang, Y., Weng, Q., 2017. Estimation of hourly and daily evapotranspiration and soil moisture using downscaled LST over various urban surfaces. *GIScience & Remote Sensing* 54 (1), 95–117. <https://doi.org/10.1080/15481603.2016.1258971>.
- Johnson, S.C., 1967. Hierarchical clustering schemes. *Psychometrika* 32 (3), 241–254.
- Joyce, R.J., Xie, P., 2011. Kalman filter-based cmorph. *Journal of Hydrometeorology* 12 (6), 1547–1563.
- Kang, S., Xu, Y., You, Q., Flügel, W.-A., Pepin, N., Yao, T., 2010. Review of climate and cryospheric change in the tibetan plateau. *Environmental research letters* 5 (1), 015101.
- Konare, A., Zakey, A., Solmon, F., Giorgi, F., Rauscher, S., Ibrah, S., Bi, X., 2008. A regional climate modeling study of the effect of desert dust on the west african monsoon. *Journal of Geophysical Research: Atmospheres* 113 (D12).
- Kumar, S.V., Dirmeyer, P.A., Peters-Lidard, C.D., Bindlish, R., Bolten, J., 2018. Information theoretic evaluation of satellite soil moisture retrievals. *Remote sensing of environment* 204, 392–400.
- Kusangaya, S., Toucher, M.L.W., van Garderen, E.A., Jewitt, G.P.W., 2016. An evaluation of how downscaled climate data represents historical precipitation characteristics beyond the means and variances. *Global Planet. Change* 144, 129–141 <https://www.sciencedirect.com/science/article/pii/S0921818116301229>.
- Lei, F., Crow, W.T., Shen, H., Su, C.-H., Holmes, T.R., Parinussa, R.M., Wang, G., 2018. Assessment of the impact of spatial heterogeneity on microwave satellite soil moisture periodic error. *Remote sensing of environment* 205, 85–99.
- Li, T., Chen, Y., Han, L., Cheng, L., Lv, Y., Fu, B., Feng, X., Wu, X., 2021. Shortened duration and reduced area of frozen soil in the northern hemisphere. *The Innovation*.
- Li, X., Al-Yaari, A., Schwank, M., Fan, L., Frappart, F., Swenson, J., Wigneron, J.-P., 2020. Compared performances of smos-ic soil moisture and vegetation optical depth retrievals based on tau-omega and two-stream microwave emission models. *Remote Sens. Environ.* 236, 111502.
- Liang, S., Fang, H., 2021. Quantitative analysis of driving factors in soil erosion using geographic detectors in Qiantang River catchment, Southeast China. *J. Soils Sediments* 21 (1), 134–147.
- Likas, A., Vlassis, N., Verbeek, J.J., 2003. The global k-means clustering algorithm. *Pattern recognition* 36 (2), 451–461.
- Liu, D., Mishra, A.K., Yu, Z., Yang, C., Konapala, G., Vu, T., 2017. Performance of SMAP, AMSR-E and LAI for weekly agricultural drought forecasting over continental United States. *J. Hydrol.* 553, 88–104 <https://www.sciencedirect.com/science/article/pii/S0022169417305140>.
- Liu, Y., Chen, Y., Wu, Z., Wang, B., Wang, S., 2021. Geographical detector-based stratified regression kriging strategy for mapping soil organic carbon with high spatial heterogeneity. *Catena* 196 (December 2019).
- Luo, P., Song, Y., Wu, P., 2021. Spatial disparities in trade-offs: economic and environmental impacts of road infrastructure on continental level. *GIScience & Remote Sensing* 58 (5), 756–775.
- Ma, H., Zeng, J., Chen, N., Zhang, X., Cosh, M.H., Wang, W., 2019. Satellite surface soil moisture from smap, smos, amsr2 and esa cci: A comprehensive assessment using global ground-based observations. *Remote Sens. Environ.* 231, 111215.
- Ma, H., Zeng, J., Zhang, X., Fu, P., Zheng, D., Wigneron, J.-P., Chen, N., Niyogi, D., 2021. Evaluation of six satellite- and model-based surface soil temperature datasets using global ground-based observations. *Remote Sens. Environ.* 264, 112605 <https://www.sciencedirect.com/science/article/pii/S00344257211003254>.
- Martínez-Fernández, J., Ceballos, A., 2005. Mean soil moisture estimation using temporal stability analysis. *J. Hydrol.* 312 (1–4), 28–38.
- McColl, K.A., Alemohammad, S.H., Akbar, R., Konings, A.G., Yueh, S., Entekhabi, D., 2017. The global distribution and dynamics of surface soil moisture. *Nat. Geosci.* 10 (2), 100–104.
- McNairn, H., Jackson, T.J., Wiseman, G., Bélair, S., Berg, A., Bullock, P., Colliander, A., Cosh, M.H., Kim, S.-B., Magagi, R., et al., 2014. The soil moisture active passive validation experiment 2012 (smapvex12): Prelaunch calibration and validation of the smap soil moisture algorithms. *IEEE Trans. Geosci. Remote Sens.* 53 (5), 2784–2801.
- Molero, B., Leroux, D., Richaume, P., Kerr, Y., Merlin, O., Cosh, M., Bindlish, R., 2018. Multi-timescale analysis of the spatial representativeness of in situ soil moisture data within satellite footprints. *Journal of Geophysical Research: Atmospheres* 123 (1), 3–21.
- Molero, B., Leroux, D.J., Richaume, P., Kerr, Y.H., Merlin, O., Cosh, M.H., Bindlish, R., 2018. Multi-Timescale Analysis of the Spatial Representativeness of In Situ Soil Moisture Data within Satellite Footprints. *Journal of Geophysical Research: Atmospheres* 123 (1), 3–21 <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017JD027478>.
- Montzka, C., Rötzer, K., Bogen, H.R., Sanchez, N., Vereecken, H., 2018. A new soil moisture downscaling approach for smap, smos, and ascats by predicting sub-grid variability. *Remote sensing* 10 (3), 427.
- Mu, Q., Zhao, M., Running, S.W., 2013. Modis global terrestrial evapotranspiration (et) product (nasa mod16a2/a3). Algorithm Theoretical Basis Document, Collection 5.
- Nielsen, D., Biggar, J., Erh, K., et al., 1973. Spatial variability of field-measured soil-water properties. *Hilgardia* 42 (7), 215–259.
- Niu, Y.J., Zhou, J.W., Yang, S.W., Wang, G.Z., Liu, L., Hua, L.M., may 2017. [Quantitative apportionment of slope aspect and altitude to soil moisture and temperature and plant distribution on alpine meadow]. *Ying yong sheng tai xue bao = The journal of applied ecology* 28 (5), 1489–1497. URL <https://doi.org/10.13287/j.1001-9332.201705.032>.
- Ochsner, T.E., Linde, E., Hafner, M., Dong, J., 2019. Mesoscale Soil Moisture Patterns Revealed Using a Sparse In Situ Network and Regression Kriging. *Water Resour. Res.* 55 (6), 4785–4800.
- Peng, J., Loew, A., Merlin, O., Verhoest, N.E., 2017. A review of spatial downscaling of satellite remotely sensed soil moisture. *Rev. Geophys.* 55 (2), 341–366.
- Peng, J., Niesel, J., Loew, A., 2015. Evaluation of soil moisture downscaling using a simple thermal-based proxy—the remedhus network (spain) example. *Hydrol. Earth Syst. Sci.* 19 (12), 4765–4782.
- Perry, M.A., Niemann, J.D., 2007. Analysis and estimation of soil moisture at the catchment scale using eofs. *J. Hydrol.* 334 (3–4), 388–404.
- Peters-Lidard, C., Zion, M., Wood, E.F., 1997. A soil-vegetation-atmosphere transfer scheme for modeling spatially variable water and energy balance processes. *Journal of Geophysical Research: Atmospheres* 102 (D4), 4303–4324.
- Planchon, O., 2000. A study of the coastal climates in france using temperature and precipitation data (1961–1990). *Meteorological Applications* 7 (3), 217–228.
- Purdy, A.J., Fisher, J.B., Goulden, M.L., Colliander, A., Halverson, G., Tu, K., Famiglietti, J.S., 2018. SMAP soil moisture improves global evapotranspiration. *Remote Sens. Environ.* 219 (September), 1–14. <https://doi.org/10.1016/j.rse.2018.09.023>.
- Qu, Y., Zhu, Z., Montzka, C., Chai, L., Liu, S., Ge, Y., Liu, J., Lu, Z., He, X., Zheng, J., et al., 2021. Inter-comparison of several soil moisture downscaling methods over the qinghai-tibet plateau, china. *J. Hydrol.* 592, 125616.
- Quinn, P., Beven, K., Culf, A., 1995. The introduction of macroscale hydrological complexity into land surface-atmosphere transfer models and the effect on planetary boundary layer development. *J. Hydrol.* 166 (3–4), 421–444.
- Redding, T.E., Hope, G.D., Fortin, M.J., Schmidt, M.G., Bailey, W.G., 2003. Spatial patterns of soil temperature and moisture across subalpine forest-clearcut edges in the southern interior of British Columbia. *Can. J. Soil Sci.* 83 (1), 121–130.
- Romshoo, S.A., 2004. Geostatistical analysis of soil moisture measurements and remotely sensed data at different spatial scales. *Environ. Geol.* 45 (3), 339–349.
- Rosenbaum, U., Bogen, H.R., Herbst, M., Huisman, J.A., Peterson, T.J., Weuthen, A., Western, A.W., Vereecken, H., 2012. Seasonal and event dynamics of spatial soil moisture patterns at the small catchment scale. *Water Resour. Res.* 48 (1).
- Sawada, Y., 2018. Quantifying Drought Propagation from Soil Moisture to Vegetation Dynamics Using a Newly Developed Ecohydrological Land Reanalysis. *Remote Sensing* 10 (8) <https://www.mdpi.com/2072-4292/10/8/1197>.

- Schaefer, G.L., Cosh, M.H., Jackson, T.J., 2007. The usda natural resources conservation service soil climate analysis network (scan). *Journal of Atmospheric and Oceanic Technology* 24 (12), 2073–2077.
- Shellito, P.J., Small, E.E., Livneh, B., 2018. Controls on surface soil drying rates observed by SMAP and simulated by the Noah land surface model. *Hydrol. Earth Syst. Sci.* 22 (3), 1649–1663.
- Silva, B.M., Silva, S.H.G., de Oliveira, G.C., Peters, P.H.C.R., dos Santos, W.J.R., Curi, N., 2014. Soil moisture assessed by digital mapping techniques and its field validation. *Ciência e Agrotecnologia* 38 (2), 140–148 http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-70542014000200005&lng=en&tng=en.
- Song, Y., Shen, Z., Wu, P., Viscarra Rossel, R.A., 2021. Wavelet geographically weighted regression for spectroscopic modelling of soil properties. *Sci. Rep.* 11, 17503 <https://doi.org/10.1038/s41598-021-96772-z>.
- Song, Y., Wang, J., Ge, Y., Xu, C., 2020. An optimal parameters-based geographical detector model enhances geographic characteristics of explanatory variables for spatial heterogeneity analysis: Cases with different types of spatial data. *GIScience & Remote Sensing* 57 (5), 593–610.
- Song, Y., Wright, G., Wu, P., Thatcher, D., McHugh, T., Li, Q., Li, S.J., Wang, X., 2018. Segment-based spatial analysis for assessing road infrastructure performance using monitoring observations and remote sensing data. *Remote Sensing* 10 (11), 1696.
- Song, Y., Wu, P., 2021. An interactive detector for spatial associations. *International Journal of Geographical Information Science* 1–26.
- Song, Y., Yang, H., Peng, J., Song, Y., Sun, Q., Li, Y., 2015. Estimating pm_{2.5} concentrations in xi'an city using a generalized additive model with multi-source monitoring data. *PLoS One* 10 (11), e0142149.
- Tao, L., Ryu, D., Western, A., Boyd, D., 2021. A New Drought Index for Soil Moisture Monitoring Based on MPDI-NDVI Trapezoid Space Using MODIS Data. *Remote Sensing* 13 (1) <https://www.mdpi.com/2072-4292/13/1/122>.
- Vereecken, H., Huisman, J.A., Pachepsky, Y., Montzka, C., van der Kruk, J., Bogena, H., Weihermüller, L., Herbst, M., Martínez, G., Vanderborght, J., 2014. On the spatio-temporal dynamics of soil moisture at the field scale. *J. Hydrol.* 516, 76–96. <https://doi.org/10.1016/j.jhydrol.2013.11.061>.
- Verhoest, N.E., Lievens, H., Wagner, W., Álvarez-Mozos, J., Moran, M.S., Mattia, F., 2008. On the soil roughness parameterization problem in soil moisture retrieval of bare surfaces from synthetic aperture radar. *Sensors* 8 (7), 4213–4248.
- Wang, J.-F., Li, X.-H., Christakos, G., Liao, Y.-L., Zhang, T., Gu, X., Zheng, X.-Y., 2010. Geographical detectors-based health risk assessment and its application in the neural tube defects study of the heshun region, china. *International Journal of Geographical Information Science* 24 (1), 107–127.
- Wang, J.-F., Zhang, T.-L., Fu, B.-J., 2016. A measure of spatial stratified heterogeneity. *Ecol. Ind.* 67, 250–256.
- Wang, T., Franz, T.E., Li, R., You, J., Shulski, M.D., Ray, C., 2017. Evaluating climate and soil effects on regional soil moisture spatial variability using eof s. *Water Resour. Res.* 53 (5), 4022–4035.
- Wang, Y., Yang, J., Chen, Y., Wang, A., De Maeyer, P., 2018. The Spatiotemporal Response of Soil Moisture to Precipitation and Temperature Changes in an Arid Region. China. *Remote Sensing* 10 (3) <https://www.mdpi.com/2072-4292/10/3/468>.
- Wei, X., Huang, C., Wei, N., Zhao, H., He, Y., Wu, X., 2019. The impact of freeze–thaw cycles and soil moisture content at freezing on runoff and soil loss. *Land Degradation & Development* 30 (5), 515–523.
- Western, A.W., Blöschl, G., Grayson, R.B., 1998. Geostatistical characterisation of soil moisture patterns in the tarrawarra catchment. *J. Hydrol.* 205 (1–2), 20–37.
- Western, A.W., Zhou, S.-L., Grayson, R.B., McMahon, T.A., Blöschl, G., Wilson, D.J., 2004. Spatial correlation of soil moisture in small catchments and its relationship to dominant spatial hydrological processes. *J. Hydrol.* 286 (1–4), 113–134.
- Wilson, D.J., Western, A.W., Grayson, R.B., 2005. A terrain and data-based method for generating the spatial distribution of soil moisture. *Adv. Water Resour.* 28 (1), 43–54.
- Wu, X., Liu, M., 2012. In-situ soil moisture sensing: measurement scheduling and estimation using compressive sensing. In: 2012 ACM/IEEE 11th International Conference on Information Processing in Sensor Networks (IPSN). IEEE, pp. 1–11.
- Xu, X., Dunbar, R., Derksen, C., Colliander, A., Kim, Y., Kimball, J., 2018. Smap l3 radiometer global and northern hemisphere daily 36 km ease-grid freeze/thaw state. NASA National Snow and Ice datacenter, Boulder, Colorado, USA.
- Yang, K., Qin, J., Zhao, L., Chen, Y., Tang, W., Han, M., Chen, Z., Lv, N., Ding, B., Wu, H., et al., 2013. A multiscale soil moisture and freeze–thaw monitoring network on the third pole. *Bull. Am. Meteorol. Soc.* 94 (12), 1907–1916.
- Zacharias, S., Bogena, H., Samaniego, L., Mauder, M., Fuß, R., Pütz, T., Frenzel, M., Schwank, M., Baessler, C., Butterbach-Bahl, K., et al., 2011. A network of terrestrial environmental observatories in germany. *Vadose zone journal* 10 (3), 955–973.
- Zappa, L., Forkel, M., Xaver, A., Dorigo, W., 2019. Deriving Field Scale Soil Moisture from Satellite Observations and Ground Measurements in a Hilly Agricultural Region. *Remote Sensing* 11 (22) <https://www.mdpi.com/2072-4292/11/22/2596>.
- Zeng, J., Chen, K.-S., Bi, H., Chen, Q., 2016. A preliminary evaluation of the smap radiometer soil moisture product over united states and europe using ground-based measurements. *IEEE Trans. Geosci. Remote Sens.* 54 (8), 4929–4940.
- Zhao, L., Yang, K., Qin, J., Chen, Y., Tang, W., Montzka, C., Wu, H., Lin, C., Han, M., Vereecken, H., 2013. Spatiotemporal analysis of soil moisture observations within a tibetan mesoscale area and its implication to regional soil moisture measurements. *J. Hydrol.* 482, 92–104.
- Zhuo, L., Dai, Q., Zhao, B., Han, D., 2020. Soil moisture sensor network design for hydrological applications. *Hydrol. Earth Syst. Sci.* 24 (5), 2577–2591 <https://hess.copernicus.org/articles/24/2577/2020/>.
- Zhuo, L., Dai, Q., Zhao, B., Han, D., 2020. Soil moisture sensor network design for hydrological applications. *Hydrol. Earth Syst. Sci.* 24 (5), 2577–2591.

A4. A locally explained heterogeneity model for examining wetland disparity

Reference: Li, Y., Luo, P., Song, Y., Zhang, L., Qu, Y. and Hou, Z., 2023. A locally explained heterogeneity model for examining wetland disparity. *International Journal of Digital Earth*, 16(2), pp.4533-4552.



A locally explained heterogeneity model for examining wetland disparity

Yang Li^{a#}, Peng Luo^{b#}, Yongze Song^c, Liqiang Zhang^a, Ying Qu^a and Zhengyang Hou^a

^aKey Laboratory of Remote Sensing Science, Faculty of Geographical Science, Beijing Normal University, Beijing, People's Republic of China; ^bChair of Cartography and Visual Analytics, Technical University of Munich, Munich, Germany; ^cSchool of Design and the Built Environment, Curtin University, Perth, Australia

ABSTRACT

Identifying the factors influencing wetland variations is crucial for understanding the relationship of climate change with wetland conservation and management. The wetland distribution is associated with multiple variables, and the interactions among these variables are complex. In this study, we aim to explore an interpretable and quantitative analysis of factors related to wetland spatiotemporal variations on the Tibetan Plateau (TP). By combining SHapley Additive exPlanations with a spatially stratified heterogeneity model, we propose a locally explained stratified heterogeneity (LESH) model that well reveals the effects of multiple variable interactions on the spatiotemporal variations of wetlands. The results show that topographic variables are the most important variables related to the spatial distribution of wetlands on the TP, and climatic variables are the most relevant factors for the increase in the wetland area on the TP from 2015 to 2019. In addition, the interactions among multiple variables strongly influence wetlands on the TP. Among them, when other geographic variables interact with the evaporation variable, its explanatory power on wetland distribution is significantly increased. Knowledge of wetland distribution determinants can help us understand the evolution of wetlands and the impacts of climate change on wetland variations.

ARTICLE HISTORY



Received 9 May 2023
Accepted 12 October 2023

KEYWORDS

Wetland distribution; spatial heterogeneity; spatial associations; SHAP

1. Introduction

Wetlands provide abundant ecological and climatic benefits. They are critical for hydrology, biogeochemical function, and biodiversity conservation (Chatterjee et al. 2015; Cohen et al. 2016; Gall et al. 2013; Russi 2013). The soil and biomass in wetlands can capture and store atmospheric carbon dioxide over long periods to counteract the effects of climate change (Chmura et al. 2003; Mitsch et al. 2013; Were et al. 2019). The Tibetan Plateau (TP) is the birthplace of many large rivers in Asia and has unique alpine wetlands (lakes, rivers, marshes, etc., under unique alpine climate conditions) (Zhao et al. 2015; Cao and Zhang 2015). As a sensitive region and magnifier of global climate change, the TP has been significantly impacted by climate conditions and environmental variability over the past three decades (Kang et al. 2007; Yao et al. 2000; Zhang et al. 2020). The spatiotemporal changes in wetlands and the relevant factors on the TP have attracted great attention.

CONTACT Liqiang Zhang  zhanglq@bnu.edu.cn  Key Laboratory of Remote Sensing Science, Faculty of Geographical Science, Beijing Normal University, Beijing, People's Republic of China

[#]These authors contributed equally to the work

© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

Knowing the factors that influence the wetland variations on the TP could facilitate our understanding of wetland development and changes. By revealing the interactions between hydrological processes and the ecosystem, we can develop coupling models to simulate spatiotemporal hydrological patterns and processes (Xue et al. 2018; Zhang et al. 2016b). These models are essential in predicting whether wetlands will be resilient or vulnerable to climate change in the future (Zhang et al. 2016a). This knowledge is also helpful for making decisions regarding wetland restoration and protection (Hu et al. 2017). To make protection measures more effective and targeted, it is essential to set different measures for different regions according to the characteristics of wetlands (Xu et al. 2019). By carefully designing these partition strategies, researchers can help to protect and preserve the unique wetland ecosystems of the TP.

Numerous methods can be used to investigate the related factors of spatiotemporal variations on wetlands. These methods can be categorized into two groups: statistical models and mechanical models (Wang et al. 2022). Correlation analysis methods are the most widely used statistical methods. The correlation of time series data was used to investigate the relationships between variables (Wang, He, and Niu 2020; Wang et al. 2022). Regression analysis methods use curve-fitting statistics to explore spatial relationships. For example, multiple regression was used to predict the wetland extent with coastal and watershed variables and calculate the explanatory power to wetland changes (Braswell and Heffernan 2019). Geographically weighted regression (GWR) was used to reveal the critical influencing factors of spatiotemporal variability on wetlands (Tian et al. 2023). In addition, several studies have investigated the impacts of climate variables on wetland changes based on mechanical models, such as wetland hydrological models (Moshir Panahi et al. 2022). Xi et al. (2020) explored the effects of temperature changes on the wetland areas across 1,250 inland Ramsar sites by estimating the wetland areal extent with a hydrological model.

Spatially stratified heterogeneity (SSH) models are effective analytical frameworks for investigating the drivers of spatial variability in geographical variables. The utilization of SSH models has been on the rise in recent years for characterizing the spatial variability of geographical variables (Guo et al. 2022; Luo et al. 2022). These models enable a comparison of the spatial distribution patterns of dependent and independent variables to calculate the power of determinants (PD) (Luo et al. 2023). A higher PD value indicates similar spatial distributions. In the classical SSH model, spatial discretization was conducted according to equal, quantile, or geometric breaks, and no optimization occurred in this process. The detected spatial associations of this method were influenced by the rule applied to determine spatial discretization. Thus, the corresponding PD could not fully explain the spatial associations between the explanatory variables and response variables. Studies have detected a significant underestimation of PD when using the classical SSH model (Luo et al. 2022). To address the above problem, several models have been developed to calculate the optimal power of determinant (OPD), such as the optimal parameter based geographical detector (OPGD) and geographically optimal zones-based heterogeneity (GOZH) models (Luo, Song, and Wu 2021; Song et al. 2020; Song and Wu 2021; Luo et al. 2022). By optimizing the spatial discretization process, the corresponding OPD can significantly improve the method to fully reveal the spatial associations.

However, current SSH models still faced difficulties to explain the contributions of variables due to the complex spatial heterogeneity of wetlands in large regions. A black box exists in the current SSH models when calculating the interactions between multiple explanatory variables. There is a need to distribute the contributions in a fair way to explain the role of each variable and the interactions between multiple variables. In addition, past studies have ignored the scale effect of geographical variables on wetlands. In spatial analyses, the size of the units directly affects the level of detail captured and the results generated (Chen et al. 2019). If the scale of the variables changes, then the covariance between them, their correlation coefficient, and the statistical model results also change (Wu 2004). Therefore, analyses relying on geographical variables are relatively scale-sensitive, and it is important to choose the optimal scale when characterizing and comparing data when using these analysis methods.

In this study, we developed a locally explained stratified heterogeneity (LESH) model to calculate the contribution of each variable for explaining the spatiotemporal heterogeneity of wetlands on the TP. The multiple grid-scale wetland density from 2000 to 2019 was calculated from a wetland product (Li et al. 2023a; Li et al. 2023b). Correspondingly, the explanatory variables were collected from Google Earth Engine (GEE) and classified into three categories: geographic, climatic, and environmental variables. The Shapley value, a commonly employed concept in cooperative game theory analyses to fairly distribute the ‘payout’ among players (Shapley 1953; Datta, Sen, and Zick 2016; Lundberg and Lee 2017), was introduced to assign a total explanatory power to each variable. Based on this model, the contribution of each variable and the interactions among multiple variables were obtained. First, the optimal scale of variable was determined by investigating the associations between each variable with wetland density at different scales. Second, the impacts of individual variables on the spatiotemporal variations of wetlands were analysed at the optimal scale. Third, the LESH model was used to calculate the interactions among multiple variables. Finally, the geographically optimal wetland zones over the first 15 years (slowly fluctuating period) and the following 5 years (rapid growth period) of the study period were determined. The impact of the related factor on the wetland distribution was analysed according to the geographically optimal zones, and each variable’s contribution was calculated.

2. Data

2.1. Response variable

In our study, the wetland density was used as the response variable. Compared with wetland area, wetland density was more convenient for comparison among different scales. Wetland density data at multiple scales were calculated from the yearly wetland product on the TP (Li et al. 2023a; Li et al. 2023b). The product was generated from Landsat satellite images between September and October from 2000 to 2019. Lakes, rivers, and marshes including moss marshes, herbaceous marshes and salt marshes were the main types of extracted wetlands. Validation experiments indicated that this wetland product is highly accurate (with a 96.1% user’s accuracy and 90.8% producer’s accuracy). Figure 1 shows the spatial (Figure 1a) and temporal (Figure 1b) variations of the wetlands on the TP. There are more wetlands in the north-western region and fewer in the south-eastern area. In terms of temporal variations, two distinct wetland periods were distinguished by Ruptures (Truong, Oudre, and Vayatis 2020), a Python library for change point detection. From 2000 to 2014, the TP wetland area fluctuated between 160,000 km² and 185,000 km² with no significant changes. From 2015 to 2019, the TP wetlands showed a rapidly extending trend. The wetland growth (50,725 km²) during this period was more than one-fourth of that in 2015. In subsequent analyses, we used wetland density from 2000 to 2014 as the proxy of the spatial variations of wetlands, and wetland density from 2015 to 2019 as the proxy of the spatial and temporal variations of wetlands.

2.2. Explanatory variables

The explanatory variables used to explain the spatial disparities of wetlands included topographical, climatic, and environmental categories derived from remotely sensed data (Table 1). All these data were collected using Google Earth Engine (GEE). Since the Landsat images used to identify wetlands were all taken from September to October, the explanatory variables were also averaged for September and October to represent the climatic and environmental conditions as much as possible.

2.2.1. Topographical variables

Topographical variables, including the elevation and derived slope data, were obtained from GEE to represent the topographical conditions of the TP. The data were collected from the digital elevation model (DEM) at a resolution of 30 m from the Space Shuttle Radar Terrain Mission (SRTM) (Farr

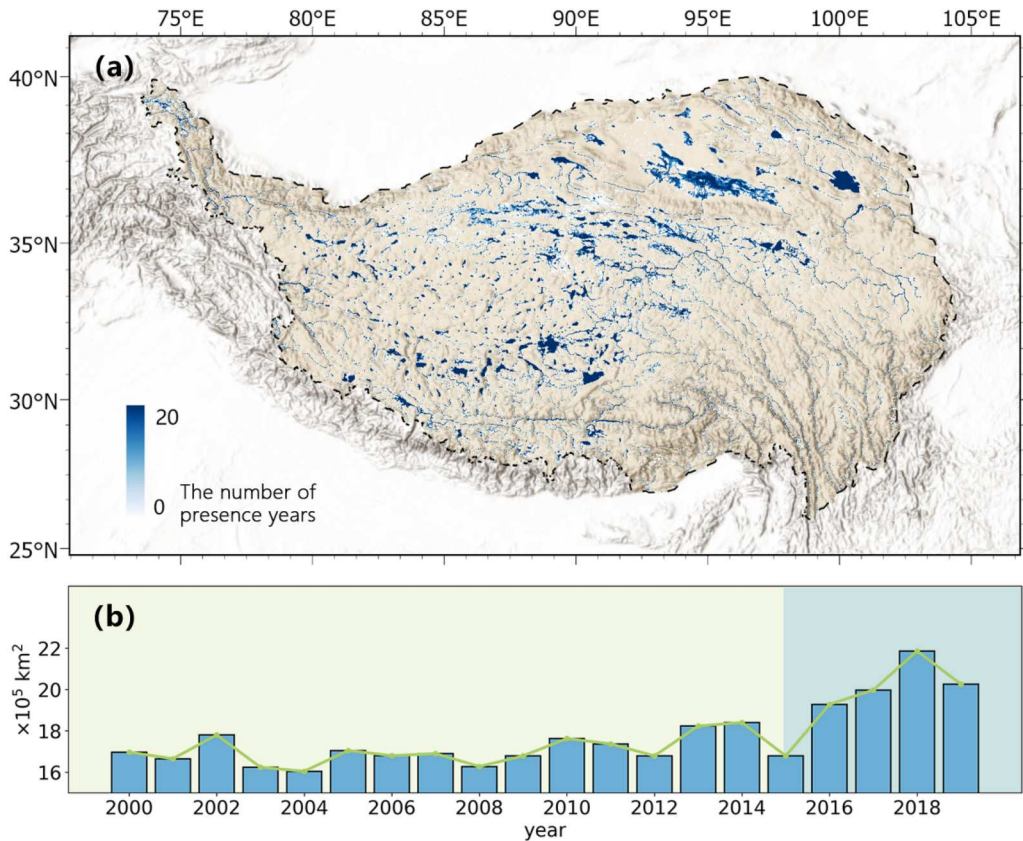


Figure 1. Distribution of TP wetlands. (a), the number of years with wetlands in each location from 2000 to 2019; (b), the TP wetland area changed from 2000 to 2019.

and Kobrick 2000). The slope data were computed from the elevation data by the spatial analysis function in GEE.

2.2.2. Climate variables

Climate change is an essential driver of the conversion of wetland ecosystems (Mitsch et al. 2013; Wang, He, and Niu 2020). The climate variables used in this study include monthly temperature and precipitation data derived from the ERA5-Land monthly averaged data (Muñoz-Sabater et al. 2021). ERA5-Land is a reanalysis product that integrates observational data with the fundamental principles of physics, thereby providing a precise characterization of the climate. The data has been transformed into monthly averages with a spatial resolution of 0.125×0.125 degrees.

Table 1. Explanatory variables for the wetland density.

Category	Variable	Abbr.	Product
Topography	Elevation	ELE	SRTM DEM
	Slope	SLO	SRTM DEM
Climate	Temperature	TEM	ERA5-Land
	Precipitation	PRE	ERA5-Land
Environment	Normalized Difference Vegetation Index	NDVI	MOD13Q1
	Enhanced Vegetation Index	EVI	MOD13Q1
	Evaporation	EVA	ERA5-Land
	Runoff	RO	ERA5-Land
	Snowmelt	SM	ERA5-Land

2.2.3. Environmental variables

The enhanced vegetation index (EVI), normalized difference vegetation index (NDVI), evaporation, runoff, and snowmelt were used as environmental variables to characterize the local environmental conditions. The EVI and NDVI data used in this study were obtained from Terra MODIS products (MOD13Q1) at a spatial resolution of 250 m (Didan, Munoz, and Huete 2015). Evaporation, runoff, and snowmelt were derived from the ERA5-Land monthly averaged data.

3. Methods

3.1. Locally explained stratified heterogeneity (LESH) model

3.1.1. Concept of the LESH

Figure 2 shows the difference between the three kinds of SSH models. Given the response variable and one or multiple explanatory variables, spatial discretization was conducted, and the variance in the response variable among the divided zones was calculated. In the classical SSH model (Figure 2b), the PD value could not fully explain the spatial associations between the explanatory variables and response variables. Some improved models, such as the OPGD and GOZH models (Figure 2c), can significantly improve the PD value to fully reveal the spatial associations. However, although the OPD can accurately estimate how multiple explanatory variables influence the response variable, we have yet to determine the contribution of each explanatory variable. For example, we may find that temperature and precipitation can together explain 50% of the wetland distribution. There is a need to determine the contribution of temperature to this interaction.

In this study, we aim to open this black box and determine the contribution of each variable to the OPD. We propose the LESH model by combining the SHAP and SSH models. The improved OPD, called the SHAP power of determinants (SPD), can fully explain the contribution of each

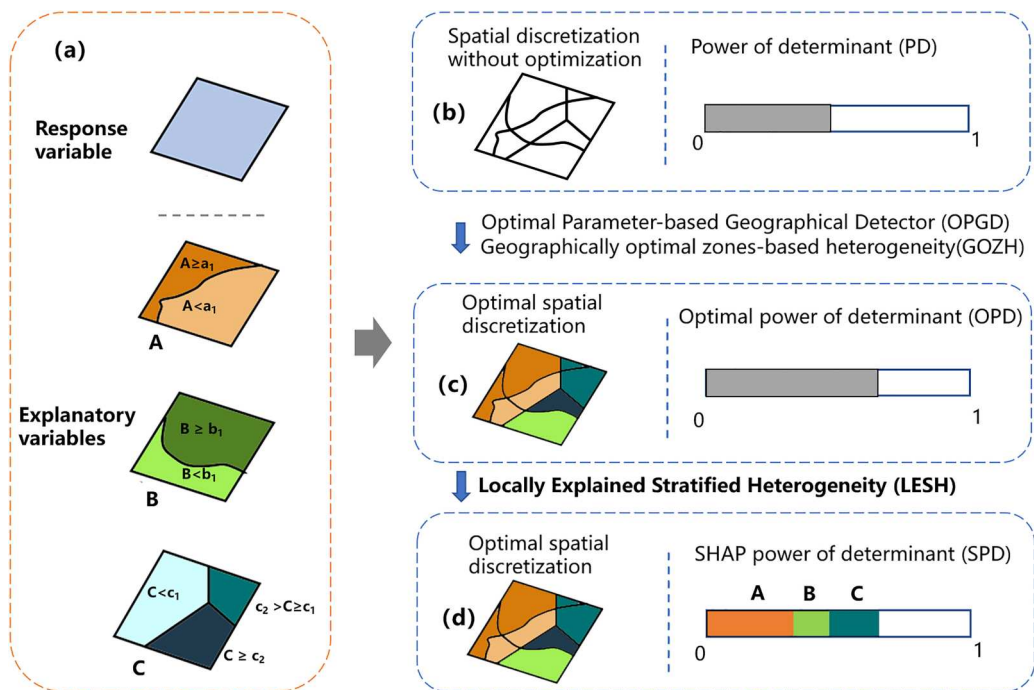


Figure 2. The development of the SSH model. (a), the response variable and the explanatory variables; (b), the concept of the power of determinant (PD); (c), the concept of the optimal power of determinant (OPD); (d), the SHAP power of determinant (SPD).

variable regardless of how many variables are considered or how complex the interaction process is (Figure 2d).

Figure 3 shows the flowchart of the LESH model. Given the response variable and multiple explanatory variables, the LESH model provides three outputs: the OPD value of individual variables and multiple variables, the SPD value of each variable, and the geographically optimal zones. The OPD values refer to the correlation between the dependent variable and explanatory variables. The SPD values represent the contribution of each variable. The geographically optimal zones can be applied to assess the overall impacts of multiple variables on spatial patterns of dependent variable.

To begin, the OPD values of individual variables and multiple variables were calculated based on the GOZH model (Figure 3b). Subsequently, the contribution of each variable was computed according to the concept of Shapley value (Figure 3d). Finally, we obtain the geographically optimal zones, which represent the strongest correlation of all possible combinations of explanatory variables to the spatial distribution of the dependent variable.

3.1.2. Power of determinants (PD) and optimal power of determinants (OPD)

The PD value was a measure of the spatial association between the response variable and the explanatory variable, with a higher PD value indicating a stronger association. This value was a ratio of the variance in the wetland density within the zones, determined by explanatory variables, to the variance across the entire study area. The formula was as follows:

$$PD = 1 - \frac{SSW}{SST} = 1 - \frac{\sum_{z=1}^h N_z \sigma_z^2}{N \sigma^2} \tag{1}$$

where *SSW* represents the summation of squares within individual zones, *SST* corresponds to the summation of squares of wetland density across the entire study area, *N_z* and *σ_z* denote the number and standard deviation of wetland density in each zone *z* (*z* = 1, ..., *h*), respectively, and *N* and *σ* are the number and standard deviation of the wetland density across the study area, respectively.

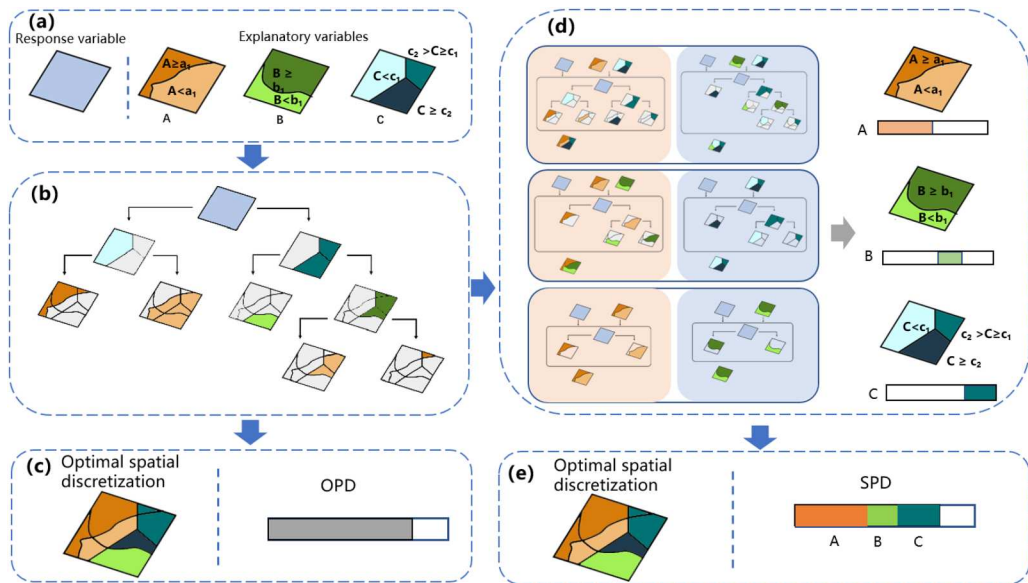


Figure 3. The workflow of the LESH model. (b), calculation of the OPD by the decision tree-based SSH model; (c), the results of OPD and the spatial discretization (zones) of the response variable; (d), calculation of the SPD value based on the SHAP model; (e), the results of SPD and the optimal spatial discretization (zones) of the response variable.

In this study, the optimal PD value, which was proposed by the geographically optimal zone-based heterogeneity (GOZH) model (Luo et al. 2022), was used to explore the factors influencing the wetland distribution. In GOZH, the OPD is represented by the Ω value, which is calculated as follows:

$$\Omega = \max(PD) = 1 - \frac{\min(SSW_{X,D})}{SSD} \quad (2)$$

where $SSW_{X,D}$ denotes the sum of squares within zones that are recorded as D and determined by explanatory variable X. The Ω value can identify the optimal geographical zone determined by multiple explanatory variables and demonstrate the maximum PD of these explanatory variables.

3.1.3. SHAP power of determinants (SPD)

The cooperative game theory proposed by Shapley (Shapley 1953; Lundberg and Lee 2017) was used to calculate the contribution of each explanatory variable under the condition of multivariate interactions. Our method can be described as follows. Suppose that the explanatory variables include $x_1, x_2, x_3, \dots, x_m$, for a total of $|M|$ (where $|M|$ represents the number of variables in the set of M) variables. $S = \{x_1, x_2, x_3, \dots, x_s\}$ ($s \leq m$) is a subset within the $|S|$ explanatory variables excluding x_j . Our method can distribute the total OPD value to each variable in a 'fair' way. For variable x_j , the SHAP power of determinants (SPD), i.e. the contribution of variable x_j to the Ω value, can be calculated by the following equation:

$$\theta_{x_j}(S) = \sum_{s \in M \setminus \{x_j\}} \frac{|S|!(|M| - |S| - 1)!}{|M|!} (\nu(S \cup \{x_j\}) - \nu(S)) \quad (3)$$

where $\theta_{x_j}(S)$ denotes the SPD of variable x_j in the set M. $S \in M \setminus \{x_j\}$ denotes that the S is a subset of M, but S does not contain the variable x_j , and $\nu(S)$ is the function used to calculate the OPD under the interaction of $|S|$ (where $|S|$ represents the number of variables in the set of S) variables.

This formula is equal to the Shapley values (Lundberg et al. 2020; Lundberg and Lee 2017) and can be understood as follows: the SPD is the weighted average of the difference between the functions $\nu(S)$ of all subsets containing the variable x_j and those not containing the variable x_j (Figure 3). Notably, the empty set is also a part of this set. In any combination of subsets, the contribution of the variable x_j can be calculated by $\nu(S \cup \{x_j\}) - \nu(S)$; then, for each variable, the mean of this contribution can be calculated over all permutations.

For a combination of variables S, the following expression can be obtained:

$$\nu(S) = \sum_{\{x_j\} \in S} \theta_{x_j}(S), \quad j = 1, 2, \dots, S \quad (4)$$

i.e. the sum of the contributions of all variables in set S is equal to the total Ω value of the set. Thus, the contribution of each variable we calculated is a part of the Ω value.

$$\theta_{x_i}(\nu(S) + \omega(S)) = \theta_{x_i}(\nu(S)) + \theta_{x_i}(\omega(S)) \quad (5)$$

The SPD proposed herein has the following three desirable properties, consistent with the Shapley values.

(1) Symmetry:

If x_i and x_j are two explanatory variables that contribute equally to all possible combinations, i.e.

$$\nu(S \cup \{x_i\}) = \nu(S \cup \{x_j\}) \quad (6)$$

for every subset S that contains neither i nor j , $s \in S \setminus \{x_j, x_i\}$, then their Shapley values are identical:

$$\theta_{x_i} = \theta_{x_j} \quad (7)$$

(2) Dummy variable:

If $v(S) = v(S \cup \{x_j\})$ for an explanatory variable x_j and all combinations $s \in M \setminus \{x_j\}$, then:

$$\theta_{x_j} = 0 \quad (8)$$

(3) Linearity:

If other methods can be used to calculate the maximum explanatory power of variables, i.e. if both $v(S)$ and $\omega(S)$ exist, then the contribution of the variable x_i in the combination of these two methods is equal to the sum of the contributions under the respective methods. This ensures the extensibility of our method.

$$\theta_{x_i}(v(S) + \omega(S)) = \theta_{x_i}(v(S)) + \theta_{x_i}(\omega(S)) \quad (9)$$

3.2. Examining wetland variations with LESH model

The framework comprises three main components (Figure 4): data pre-processing, optimal scale determination, and the computational stage utilizing the LESH model. This stage involves individual variable exploration, multiple variables exploration and identifying optimal zones.

3.2.1. Data pre-processing and the identification of the optimal analysis scale

The raw wetland data used herein were pixel-based classification images, but the LESH model requires continuous variables, such as area and density variables. Therefore, the data had to be transformed into the wetland density, i.e. the proportion of wetlands per unit area. To investigate the effects of the variable scale on the spatial distribution of wetlands, the multi-resolution data was aggregated using average function. The wetland density was aggregated from the 1-km resolution to the 10-km resolution at 1-km intervals and from the 10-km resolution to the 150-km resolution at 10-km intervals. The explanatory variables were also aggregated to different resolutions using the average function. For the elevation, slope, NDVI, and EVI variables, we first obtained 1-km-resolution data through the GEE platform and then obtained multi-resolution data by calculating the pixel means. For the five variables of temperature, precipitation, evaporation, runoff, and snowmelt, since the original resolution was $0.125^\circ \times 0.125^\circ$, we first resampled these data to a 1-km resolution using the GEE platform and then aggregated them to other resolutions by calculating the average values.

In this study, we investigate the optimal scale to analyse the distribution of wetlands on the TP using multiscale data in the LESH model. First, for each year and each scale, the OPD values of all nine variables were calculated. Second, for each scale, a box plot of all OPD values of the nine variables over two decades was obtained. Third, the mean OPD values of the nine variables over two decades were calculated. Finally, the optimal scale was selected according to the change rate between the adjacent scales. The locally estimated scatterplot smoothing (LOESS) model (Jacoby 2000) was used to fit the mean OPD values into a curve, and then the change rate of OPD values at different scales was calculated. The scale with a change rate of less than 5% was selected as the optimal scale (Song et al. 2020; Luo, Song, and Wu 2021).

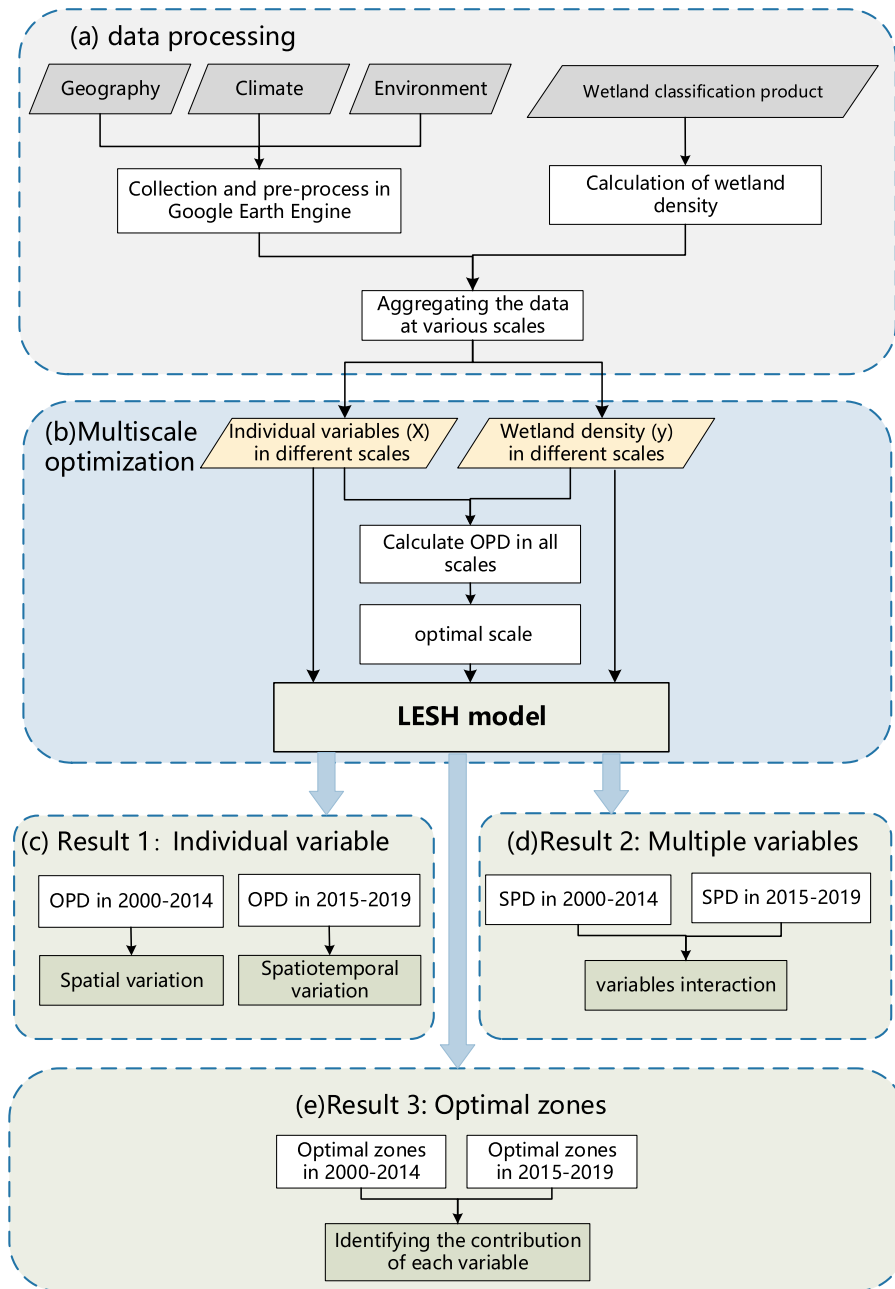


Figure 4. Schematic workflow of the process used to identify the spatiotemporal heterogeneity and influencing factors of the wetland distribution on the TP based on the LESH model.

3.2.2. Calculating of OPD values of individual variables

Based on the optimal scale, we analysed the effects of individual explanatory variables on the wetland density from 2000 to 2019 using the LESH model. First, the annual data were leveraged to calculate the OPD values. The order of importance of the variables was determined by comparing the OPD values calculated from the individual variables. Second, the annual wetland density and explanatory variables were classified into two periods considering that the wetland area showed

two distinct phases, i.e. 2000-2014 and 2015-2019. The change rates of each explanatory variable in both periods were calculated. By comparing the changes in OPD values of each explanatory variable, we obtained the key variables dominating the wetland distributions in different periods.

3.2.3. Calculating of SPD values

This study detected and analysed the interactions of two variables as a case study. We calculated the interactions between variable pairs and their respective contributions (SPD) using the LESH model. Specifically, the annual data were classified into two periods. Then, we iterated through all possible combinations of variables to compute the OPD under different combinations of variables. Finally, the SPD values were calculated for each variable. The SPD value is the weighted average of all OPD values of the combinations containing the targeted variable with that of the combinations without the targeted variable.

3.2.4. Identification of geographically optimal zones

Stratified variables from the spatial discretization were used to identify the geographically optimal zones. According to the stratified variables, wetland density was grouped into geographically optimal zones, indicating the highest homogeneity within zones and the highest heterogeneity between zones.

Since all possible combinations of variables were iterated in the above calculations, we can obtain the geographically optimal combination of variables that is most relevant to wetland density. Subsequently, the contribution of each variable to these optimal zones was analysed, thereby providing insights into how multiple variables impact the spatial patterns of wetlands. Finally, a comprehensive analysis of the geographical, climate, and environmental variables was conducted based on the geographically optimal zones, aiming to reveal the regional spatial associations with the wetland density.

4. Results

4.1. The optimal scale for analysing wetland distribution

The optimal scale was identified for the heterogeneity analysis of the wetland distribution. **Figure 5a** shows the variations in Ω values at different scales. The explanatory power of the wetland distribution gradually increased with the scale of analysis. From the 1-km to the 60-km scale, the growth rate consistently increased. The highest value (0.140) was reached at the 60-km scale (**Figure 5b**). As the scale increased, the growth rate of Ω value started to decelerate. By the scale reached 120-km, the growth rate of Ω value was lower than 0.05. The choice of scale involved a trade-off between the strength of interpretation and the granularity of the study, since a high

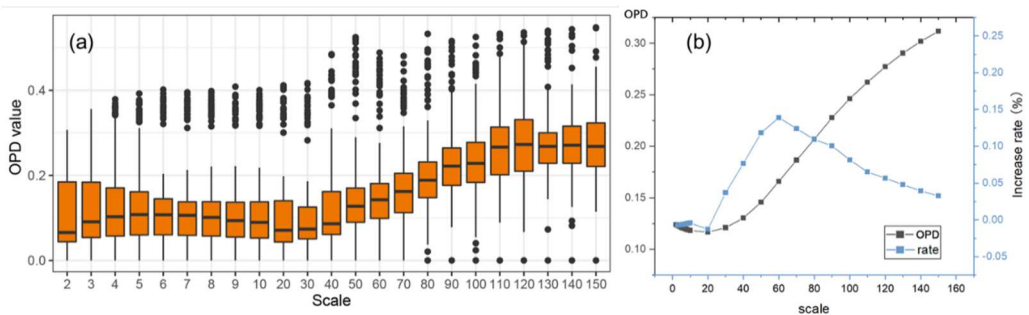


Figure 5. Selection of the optimal scale for the wetland distribution analysis. (a), the OPD (Ω) values calculated at different scales, with each box containing 20 years of results. (b), the fitting curve of OPD based on LOESS model (black) and the growth rates of OPD values (blue).

scale leads to a decrease in the number of image elements and a decrease in the granularity of the study. In this study, we chose 120-km, where the growth rate of Ω value was below 0.05, as the optimal scale for analysis.

4.2. Impacts of individual variables

Figure 6 shows the Ω values of different variables at the 120-km scale over 20 years. From the perspective of the three categories of variables, topographic variables had the highest spatial associations with wetland density, followed by climatic and environmental variables. In terms of individual variables, the slope variable has the highest average Ω value among the nine variables, explaining 49.2% of the spatial variability in the wetland distribution on average (20 years). Among the nontopographic variables, vegetation accounted for the most important role in the wetland distribution. Among them, the Ω values of EVI and NDVI were 0.320 and 0.318, respectively. Runoff had a Ω value of 0.279 for the wetland distribution.

We further examined the spatial associations between the individual variables and the wetland density during two periods. Figure 7 shows the Ω values of each variable during 2000-2014 and 2015-2019. The results also reveal the dominant impact of the topographical variables on the wetland distribution. However, the importance of variables changed during the two periods. During 2015-2019, the Ω values of topographical variables such as the slope and elevation decreased by 17.6% and 26.0%, respectively, while the Ω values of NDVI, EVI and temperature increased by 26.0%, 43.1%, and 73.5% (Figure 7b), respectively, compared with 2000-2014. The decreased Ω values of the topographical variables indicated a decrease in the explanatory power on the wetland distribution. This suggested that the newly formed wetlands had fewer spatial associations with the topographical variables but were more significantly influenced by meteorological variables such as temperature and environmental variables.

4.3. Impacts of the interactions of variable pairs

The proposed LESH model can reveal the contribution of each variable in the interactions of multiple variables. The results show that the interactions between topographic variables and other variables had the strongest explanatory power for the wetland distribution (Figure 8). Environmental and climatic variables play a secondary role in explaining the distribution of wetlands. Among the interactions involving nontopographic variables, the interactive effect of NDVI and TEM had the greatest impact on the wetland distribution. The Ω values were 0.40 and 0.43 in the 2000-2014 and 2015-2019 periods, respectively.

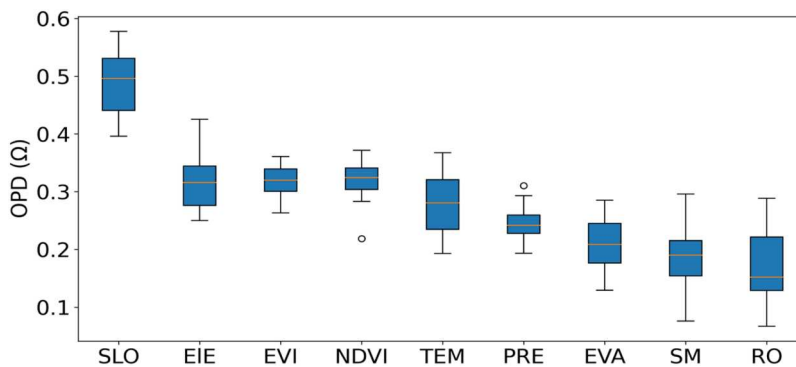


Figure 6. The OPD values of the explanatory variables of the wetland distribution. SLO refers to slope, EIE refers to elevation, EVI refers to enhanced vegetation index, NDVI refers to Normalized Difference Vegetation Index, TEM refers to temperature, PRE refers to precipitation, EVA refers to Evaporation, SM refers to snowmelt, RO refers to runoff.

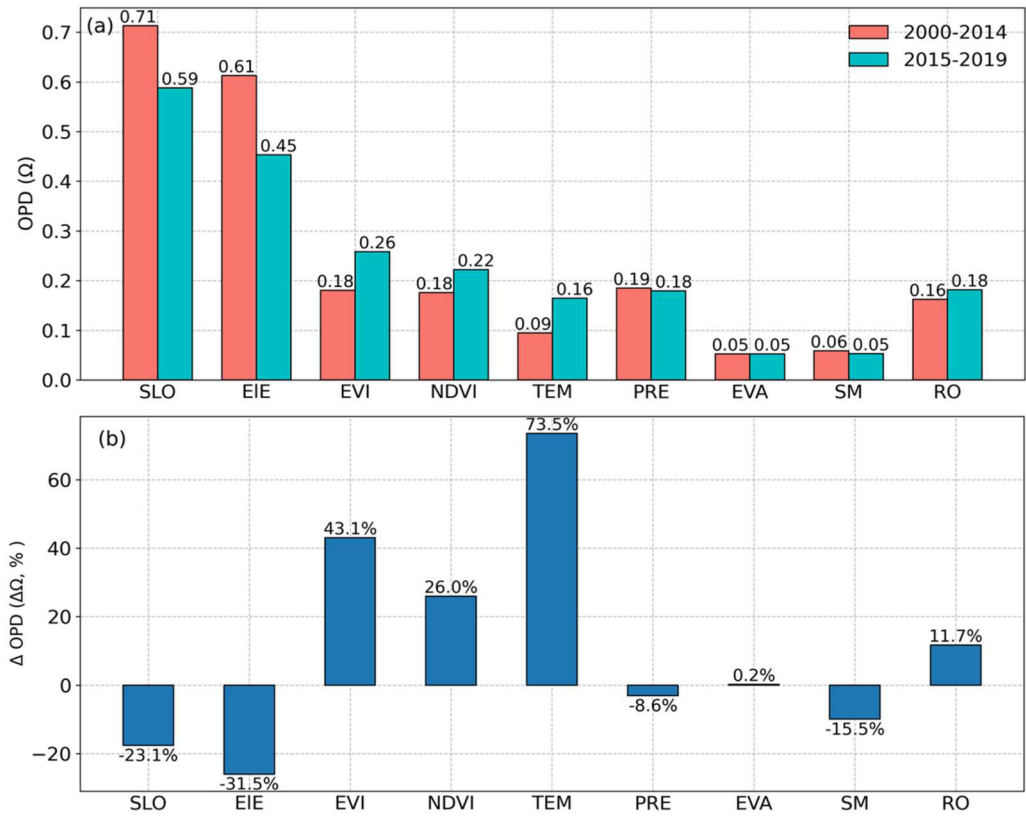


Figure 7. The change of OPD values in the two periods. (a), the OPD values in the 2000-2014 and 2015-2019; (b), the percentage change of OPD values between the two periods. SLO refers to slope, ELE refers to elevation, EVI refers to enhanced vegetation index, NDVI refers to Normalized Difference Vegetation Index, TEM refers to temperature, PRE refers to precipitation, EVA refers to Evaporation, SM refers to snowmelt, RO refers to runoff.

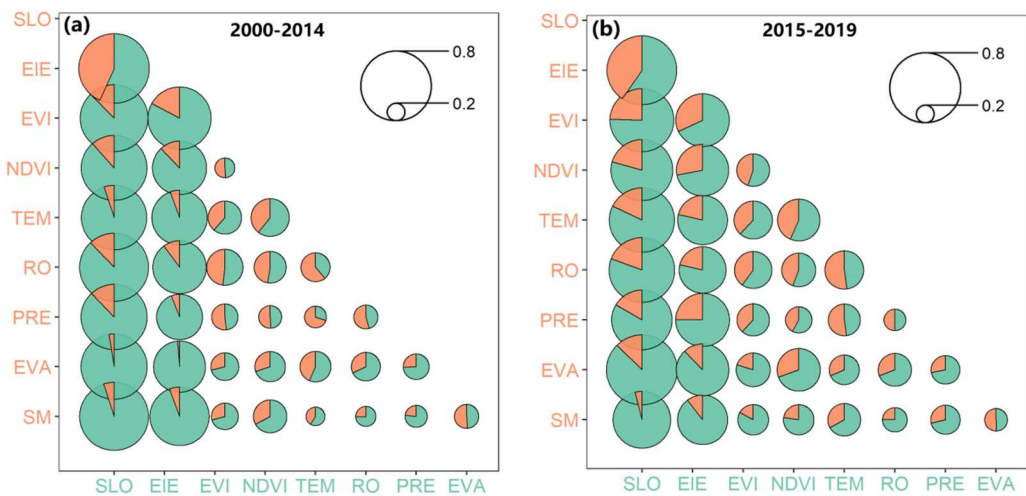


Figure 8. SHAP power of determinants (SPD) values between two variables. (a), in the 2000-2014; (b), in the 2015-2019. The pie chart proportions correspond to variables on the x- and y-axes and are distinguished by colors. SLO refers to slope, ELE refers to elevation, EVI refers to enhanced vegetation index, NDVI refers to Normalized Difference Vegetation Index, TEM refers to temperature, PRE refers to precipitation, EVA refers to Evaporation, SM refers to snowmelt, RO refers to runoff.

It is noteworthy that some nonlinearly enhanced interactions between the variables were detected, as evidenced by the interactive Ω value being greater than the sum of the individual Ω values of the two variables. For example, the interactions between evaporation and other variables exhibited nonlinearly enhanced effects. In the 2000-2014 period, the Ω value of evaporation and temperature was 0.325, of which evaporation accounted for 0.141 and temperature accounted for 0.184. When the individual variables were used, the Ω value of temperature was 0.09 and that of evaporation was 0.05. In the 2015-2019 period, the Ω value of evaporation and slope was 0.71 (0.62 for slope and 0.09 for evaporation). When the individual variables were used, the Ω value of the slope was 0.59 and that of evaporation was 0.05. In addition, we found that the higher the correlation between variables was, the weaker the interaction-derived enhancement was. For example, NDVI and EVI exhibited high correlations, with correlation coefficients of 0.98, while the interaction between these two variables was very weak.

For the temporal pattern, the topographic variables (e.g. elevation, slope) accounted for significantly lower proportions of the pairwise interactions between variables in the 2015-2019 period than in the 2000-2014 period (Figure 8). This finding was consistent with the results of the individual variables analysis, i.e. increased wetlands were not significantly correlated with the topographic variables.

4.4. Optimal variable combinations and heterogeneous spatial partitioning

The geographically optimal wetland distribution zones in the two periods were detected by the LESH model. As shown in Figure 9, in the 2000-2014 period, four variables, the slope, elevation, runoff, and precipitation, were used to identify the 12 wetland zones on the TP. All zones could be divided into two groups based on slope. The first group included zones A and B with slopes greater than 5.5° . These zones were located mainly in the south-eastern region and along the margins of the TP, where a large number of high mountains are distributed, resulting in steep gradients at large scales (120 km). The wetlands here are predominantly low-density riverine wetlands. The

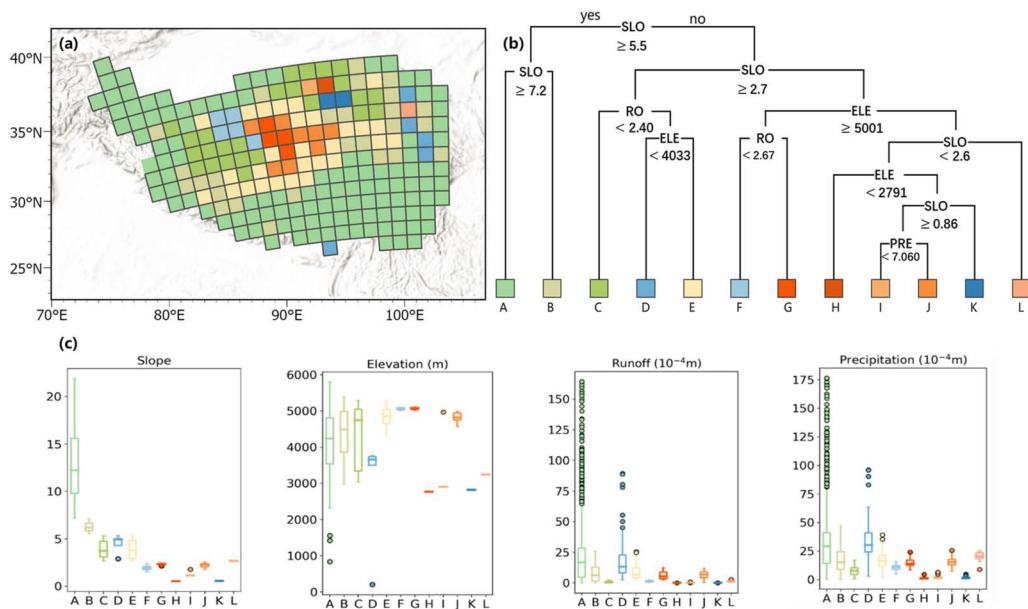


Figure 9. Geographically optimal results of the LESH model. (a), Geographically optimal wetland zones on the TP in 2000-2014; (b), the decision tree of identifying optimal zones; (c), and statistical summaries of explanatory variables in each optimal zone. SLO refers to slope, ELE refers to elevation, PRE refers to precipitation, RO refers to runoff.

second group was composed of the remaining zones, which had slopes less than 5.5°. These zones were located mainly in the central TP, that is, the source region of the three rivers and the Qiangtang Plateau. Plateaus dominate this region with small changes in slope, and the factors that control the distribution of wetlands include multiple variables.

During the period from 2015 to 2019, wetlands on the TP were grouped into 11 regions by five variables: the slope, elevation, EVI, temperature, and NDVI (Figure 10). The rules for dividing the zones in 2015-2019 differed from those applied in 2000-2014. For example, the region that was divided into regions E and I in 2015-2019 was divided into five regions (C, E, F, G, J) in 2000-2004, with a more fragmented distribution. Climate change on the TP has also led to changes in division rules even though the same regions, for example, the division rules between region K and the other regions were dominated by the slope until 2014 but became temperature-dependent from 2015-2019.

Figure 11 shows the SPD values of the optimal variable combinations that determine the geographically optimal zones in the two periods. In 2000-2014, the interactions of four variables explained 77.3% of the wetland distribution on the TP, a significantly greater proportion than that explained by single variables or when using all nine variables. The slope was the most dominant variable, contributing 41% of the explanatory power, and the elevation variable contributed 25%. Runoff and precipitation contributed the remaining 10% of the explanatory power. From 2000 to 2014, the wetland area did not change extensively, so static variables such as the slope and elevation contributed largely to the wetland distribution.

During the 2015-2019 period, the five variables of the slope, elevation, EVI, temperature, and NDVI divided the TP into 11 regions. The interactions among these five variables explained 75.1% of the wetland distribution on the TP. Unlike the previous 15 years, the explanatory powers of the slope and elevation on wetland distribution decreased to 33% and 19%, respectively. The EVI and temperature became the main explanatory factors during the period of rapid wetland growth, contributing 9% and 7%, respectively.

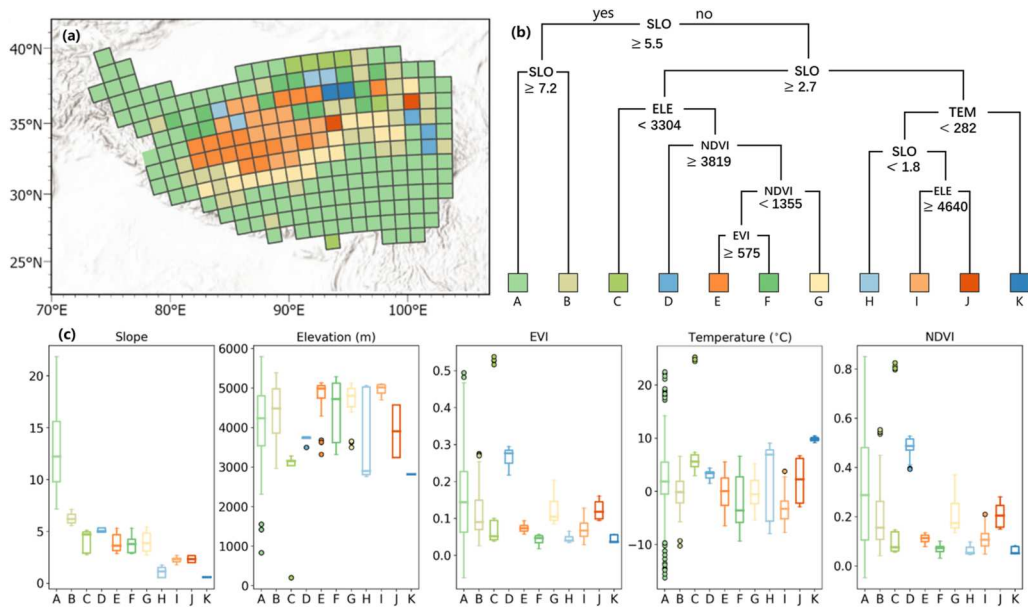


Figure 10. Geographically optimal results of the LESH model. (a), Geographically optimal wetland zones on the TP in 2000-2014; (b), the decision tree of identifying optimal zones; (c), the statistical summaries of explanatory variables in each optimal zone. SLO refers to slope, ELE refers to elevation, EVI refers to enhanced vegetation index, NDVI refers to Normalized Difference Vegetation Index, TEM refers to temperature.

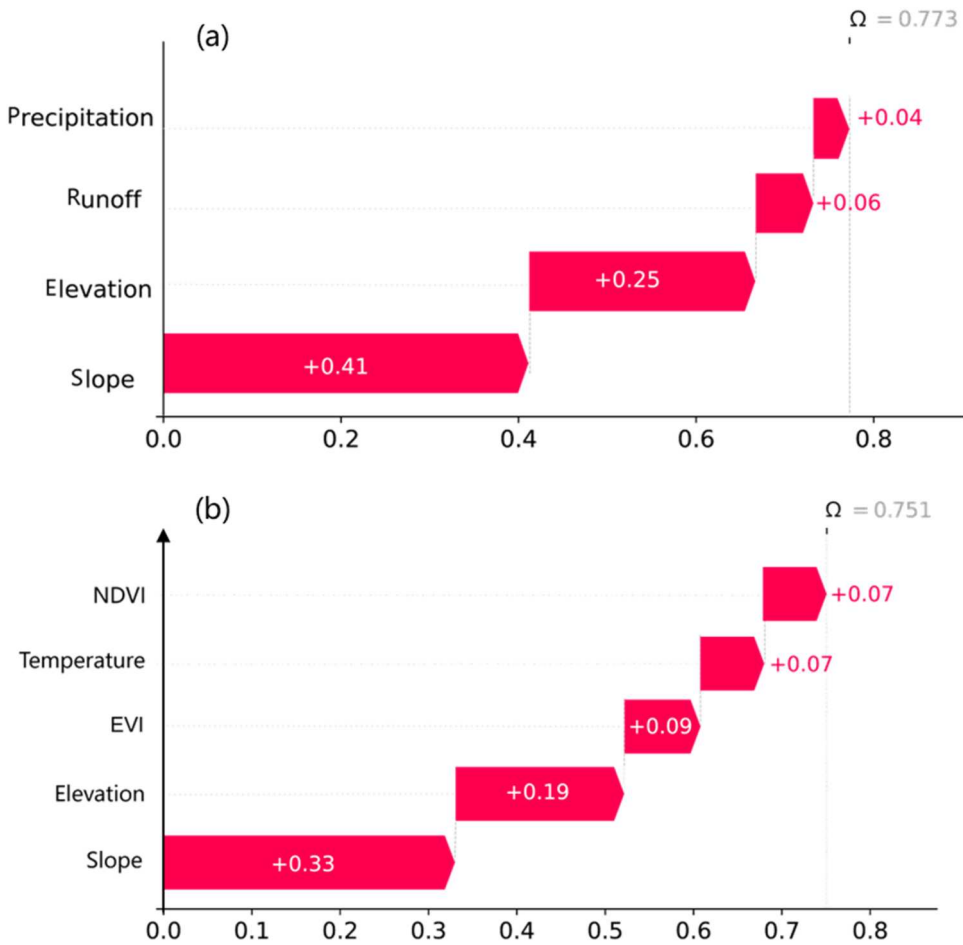


Figure 11. The SPD values of key variables. (a) shows the SPD values in 2000-2014; (b) shows the SPD values in 2015-2019.

5. Discussion

5.1. Methodological contributions

This study proposed a LESH model to investigate the linkages between the wetland distribution and geographical variables on the TP. Compared with other methods (Table 2), the LESH model has the following advantages.

First, the LESH model can accurately calculate the spatial associations between geographical variables and the wetland distribution, including linear and nonlinear relationships. However, linear regression models cannot capture the nonlinear relationships between variables.

Table 2. Comparison of several commonly used methods.

Method feature	CA	MLR	GWR	GOZH	LESH
Nonlinear relationship detection				Y	Y
Interaction relationship detection				Y	Y
Contribution assignment		Y	Y		Y
Interpretability	Y	Y	Y	Y	Y
Local spatial variations			Y		

CA: Correlation analysis (e.g. Pearson Correlation Coefficient), MLR: Multivariable linear regression, GWR: Geographically weighted regression, GOZH: Geographically optimal zones-based heterogeneity.

Second, the LESH model explores spatial associations by considering the interaction among explanatory variables. Geographical variables often exhibit complex interrelationships and are seldom independent of each other (Song and Wu 2021). Therefore, for linear models, non-independent variables can result in unstable outcomes, reduced explanatory power, and even erroneous conclusions (Brauer and Curtin 2018).

Third, the LESH model can fairly allocate the contributions of each variable to the spatiotemporal distribution of wetlands. SPD value provides a consistent and objective approach to discerning the variable with the most substantial influence. The SPD value based on Shapley value is the only method of contribution assignment that can satisfy several desirable theoretical properties.

Finally, the LESH model is interpretable throughout the process, including decision tree-based OPD calculation and optimal zones identification, and variable contribution assignment based on Shapley value. It should be mentioned that GWR is the only algorithm among those compared that can map out the spatial correlation within local regions (Local spatial variations).

The LESH model can also be applied to study similar problems, especially those involving complex interactions among multiple variables. The LESH model is desirable for calculating the contribution of each variable and the interactions between variables and can be used in factor analyses, driving force explorations, and other applications in different fields, such as the natural sciences, social sciences, and environmental pollution (Wang and Xu 2017; Guo et al. 2022).

5.2. Limitations and interpretations of the findings

The proposed LESH model investigated the associations between the response variable and explanatory variables based on statistical data analysis rather than by inferring causality from the mechanism. Based on the LESH model, the related factors of the wetland spatiotemporal variation were detected. We obtained four main findings:

First, 120-km was found to be a suitable scale for exploring the relationships between the wetland distribution and environmental variables on the TP. Under this condition, the spatial associations between wetlands and environmental variables are highly significant.

Second, we found that topographic variables were the most important variables determining the spatial distribution of wetlands on the TP, while climate variables were important in controlling the increased wetland areas. Temperature had an essential effect on the wetland area changes. Over the past few decades, the TP has generally become increasingly warm and wet (Kuang and Jiao 2016). As a region sensitive to global climate change, the TP is warming more rapidly than the worldwide average (Duan and Xiao 2015). The period from 2015 to 2019 was the hottest five-year period on record (Global Climate Status Statement 2019). The increased glacier meltwater and earlier permafrost thawing due to global warming have provided more abundant water recharge, resulting in the formation of new wetlands. The wetland data and explanatory variables we used were both derived from the mean values of September-October, which might introduce a bias in the result. The TP receives most of its precipitation in summer (Zhu and Sang 2018), so a greater influence of temperature and precipitation on wetlands would be found using summer data. Nonetheless, considering that we used multi-year data and focused on the interannual variation of wetlands, this mitigates the uncertainty caused by the data time of September-October.

Third, we investigated the interactions between two variables and the contribution of each variable. The spatiotemporal variabilities in wetlands are influenced by complex geographical factors and their interactions. The results show that the interactions of topographic variables with other variables have the strongest explanatory powers for the wetland distribution. Among the interactions involving nontopographic variables, the synergistic effect of the NDVI and TEM variables had the greatest influence on wetland distribution. In addition, nonlinear enhancement effects were observed between evaporation and several other variables, such as between evaporation and temperature, evaporation and precipitation, and evaporation and NDVI. Although there is a complex

feedback mechanism between wetlands and vegetation, vegetation is more likely to respond to wetland changes rather than being the driver of wetland spatiotemporal variabilities.

Finally, we identified the geographically optimal wetland distribution zones in two periods using the LESH model. We identified the major factors influencing wetlands in different regions within different periods, thereby enhancing our knowledge and understanding of the drivers of the spatiotemporal pattern of TP wetlands.

Our study also has the following limitation: the impact of human activities is not considered due to a lack of quantifiable data on the entire study area. Previous research shows that agricultural activities are an important driver in the decline of wetlands (Nie and Li 2011). However, compared to the climate change on the TP wetlands, the impact of human activities may be very limited (Chen et al. 2013).

6. Conclusion

In this study, a locally explained heterogeneity model was proposed to explore the heterogeneity of the spatiotemporal distribution of wetlands on the TP from 2000 to 2019. The LESH model can reveal the maximum spatial associations between the wetland density and multiple related variables and can fairly distribute the contributions of each variable and the interactions among multiple variables. Based on this model, we sought to improve our general understanding of how (and which) spatial factors are related to the wetland distribution and the extent to which the variation in the wetland distribution across the TP can be explained by geographical factors.

The results of the spatial patterns of wetland density variabilities in different phases obtained with the LESH model show that topographic variables (slope and elevation) were the most important variables determining the spatial distribution of wetlands on the TP, and temperature was an important reason for the increased wetland area observed from 2015 to 2019 on the TP. Multivariate interactions increased the explanatory power of the model to wetland distribution on the TP, and the interactions of the evaporation variable with other variables had enhancement effects.

This study enriches the theory of spatial stratification heterogeneity and analyses the wetland distribution heterogeneity in the TP over the past 20 years. Knowledge of wetland distribution determinants can help us understand the development and evolution of wetlands and the impacts of climate change on wetlands. The results of optimal wetland zoning can comprehensively reflect the regional natural geographic characteristics, provide a basis for the regional division of the TP, and serve biodiversity protection and nature reserve construction.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This project was supported by the National Natural Science Foundation of China (Grant No. 41925006).

Data availability statement

The data and code used in this study dataset are available by contacting the corresponding author.

References

Braswell, Anna E., and James B. Heffernan. 2019. "Coastal Wetland Distributions: Delineating Domains of Macroscale Drivers and Local Feedbacks." *Ecosystems* 22 (6): 1256–1270. <https://doi.org/10.1007/s10021-018-0332-3>.

- Brauer, Markus, and John J. Curtin. 2018. "Linear Mixed-Effects Models and the Analysis of Nonindependent Data: A Unified Framework to Analyze Categorical and Continuous Independent Variables That Vary Within-Subjects and/or Within-Items." *Psychological Methods* 23 (3): 389–411. <https://doi.org/10.1037/met0000159>.
- Cao, Shixiong, and Junze Zhang. 2015. "Political Risks Arising from the Impacts of Large-Scale Afforestation on Water Resources of the Tibetan Plateau." *Gondwana Research* 28 (2): 898–903. <https://doi.org/10.1016/j.gr.2014.07.002>.
- Chatterjee, Kajal, Abhirup Bandyopadhyay, Amitava Ghosh, and Samarjit Kar. 2015. "Assessment of Environmental Factors Causing Wetland Degradation, Using Fuzzy Analytic Network Process: A Case Study on Keoladeo National Park, India." *Ecological Modelling* 316: 1–13. <https://doi.org/10.1016/j.ecolmodel.2015.07.029>.
- Chen, Lei, Yong Gao, Di Zhu, Yihong Yuan, and Yu Liu. 2019. "Quantifying the Scale Effect in Geospatial Big Data Using Semi-Variograms." *PLoS ONE* 14 (11): e0225139.
- Chen, Huai, Qiuan Zhu, Changhui Peng, Ning Wu, Yanfen Wang, Xiuqing Fang, Yongheng Gao, et al. 2013. "The Impacts of Climate Change and Human Activities on Biogeochemical Cycles on the Qinghai-Tibetan Plateau." *Global Change Biology* 19 (10): 2940–2955. <https://doi.org/10.1111/gcb.12277>.
- Chmura, Gail L., Shimon C. Anisfeld, Donald R. Cahoon, and James C. Lynch. 2003. "Global Carbon Sequestration in Tidal, Saline Wetland Soils." *Global Biogeochemical Cycles* 17 (4): 1111.
- Cohen, Matthew J, Irena F Creed, Laurie Alexander, Nandita B Basu, Aram JK Calhoun, Christopher Craft, Ellen D'Amico, et al. 2016. "Do Geographically Isolated Wetlands Influence Landscape Functions?" *Proceedings of the National Academy of Sciences* 113 (8): 1978–1986. <https://doi.org/10.1073/pnas.1512650113>.
- Datta, Anupam, Shayak Sen, and Yair Zick. 2016. "Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems." In *2016 IEEE Symposium on Security and Privacy (SP)*, 598–617. IEEE.
- Didan, Kamel, Armando Barreto Munoz, and Alfredo Huete. 2015. "MODIS Vegetation Index User's Guide (MOD13 Series)."
- Duan, Anmin, and Zhixiang Xiao. 2015. "Does the Climate Warming Hiatus Exist Over the Tibetan Plateau?" *Scientific Reports* 5 (1): 13711. <https://doi.org/10.1038/srep13711>.
- Farr, Tom G., and Mike Kobrick. 2000. "Shuttle Radar Topography Mission Produces a Wealth of Data." *Eos, Transactions American Geophysical Union* 81 (48): 583–585. <https://doi.org/10.1029/EO081i048p00583>.
- Gall, Heather E, Jeryang Park, Ciaran J Harman, James W Jawitz, P. Suresh, and C. Rao. 2013. "Landscape Filtering of Hydrologic and Biogeochemical Responses in Managed Catchments." *Landscape Ecology* 28 (4): 651–664. <https://doi.org/10.1007/s10980-012-9829-x>.
- Guo, Jiangang, Jinfeng Wang, Chengdong Xu, and Yongze Song. 2022. "Modeling of Spatial Stratified Heterogeneity." *GIScience & Remote Sensing* 59 (1): 1660–1677. <https://doi.org/10.1080/15481603.2022.2126375>.
- Hu, Shengjie, Zhenguo Niu, Yanfen Chen, Lifeng Li, and Haiying Zhang. 2017. "Global Wetlands: Potential Distribution, Wetland Loss, and Status." *Science of The Total Environment* 586 (May): 319–327. <https://doi.org/10.1016/j.scitotenv.2017.02.001>.
- Jacoby, William G. 2000. "Loess: A Nonparametric, Graphical Tool for Depicting Relationships between Variables." *Electoral Studies* 19 (4): 577–613. [https://doi.org/10.1016/S0261-3794\(99\)00028-1](https://doi.org/10.1016/S0261-3794(99)00028-1).
- Kang, ShiChang, YongJun Zhang, DaHe Qin, JiaWen Ren, QiangGong Zhang, Bjorn Grigholm, and Paul A. Mayewski. 2007. "Recent Temperature Increase Recorded in an Ice Core in the Source Region of Yangtze River." *Chinese Science Bulletin* 52 (6): 825–831. <https://doi.org/10.1007/s11434-007-0140-1>.
- Kuang, Xingxing, and Jiu Jimmy Jiao. 2016. "Review on Climate Change on the Tibetan Plateau During the Last Half Century." *Journal of Geophysical Research: Atmospheres* 121 (8): 3979–4007. <https://doi.org/10.1002/2015JD024728>.
- Li, Yang, Zhengyang Hou, Liqiang Zhang, Ying Qu, Guoqing Zhou, Jintai Lin, Jingwen Li, and Ke Huang. 2023a. "Long-Term Spatio-Temporal Changes of Wetlands in Tibetan Plateau and Their Response to Climate Change." *International Journal of Applied Earth Observation and Geoinformation* 121 (July): 103351. <https://doi.org/10.1016/j.jag.2023.103351>.
- Li, Yang, Zhengyang Hou, Liqiang Zhang, Changqing Song, Shilong Piao, Jintai Lin, Shushi Peng, et al. 2023b. "Rapid Expansion of Wetlands on the Central Tibetan Plateau by Global Warming and El Niño." *Science Bulletin* 68 (5): 485–488. <https://doi.org/10.1016/j.scib.2023.02.021>.
- Lundberg, Scott M., Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. "From Local Explanations to Global Understanding with Explainable AI for Trees." *Nature Machine Intelligence* 2 (1): 56–67. <https://doi.org/10.1038/s42256-019-0138-9>.
- Lundberg, Scott M, and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In *Advances in Neural Information Processing Systems* 30: 4765–4774.
- Luo, Peng, Yongze Song, Xin Huang, Hongliang Ma, Jin Liu, Yao Yao, and Liqiu Meng. 2022. "Identifying Determinants of Spatio-Temporal Disparities in Soil Moisture of the Northern Hemisphere Using a Geographically Optimal Zones-Based Heterogeneity Model." *ISPRS Journal of Photogrammetry and Remote Sensing* 185 (March): 111–128. <https://doi.org/10.1016/j.isprsjprs.2022.01.009>.

- Luo, Peng, Yongze Song, and Peng Wu. 2021. "Spatial Disparities in Trade-Offs: Economic and Environmental Impacts of Road Infrastructure on Continental Level." *GIScience & Remote Sensing* 58 (5): 756–775. <https://doi.org/10.1080/15481603.2021.1947624>.
- Luo, Peng, Yongze Song, Di Zhu, Junyi Cheng, and Liqiu Meng. 2023. "A Generalized Heterogeneity Model for Spatial Interpolation." *International Journal of Geographical Information Science* 37 (3): 634–659. <https://doi.org/10.1080/13658816.2022.2147530>.
- Mitsch, William J., Blanca Bernal, Amanda M. Nahlik, Ülo Mander, Li Zhang, Christopher J. Anderson, Sven E. Jørgensen, and Hans Brix. 2013. "Wetlands, Carbon, and Climate Change." *Landscape Ecology* 28 (4): 583–597. <https://doi.org/10.1007/s10980-012-9758-8>.
- Moshir Panahi, Davood, Georgia Destouni, Zahra Kalantari, and Bagher Zahabiyou. 2022. "Distinction of Driver Contributions to Wetland Decline and Their Associated Basin Hydrology Around Iran." *Journal of Hydrology: Regional Studies* 42 (August): 101126.
- Muñoz-Sabater, Joaquín, Emanuel Dutra, Anna Agustí-Panareda, Clément Albergel, Gabriele Arduini, Gianpaolo Balsamo, Souhail Boussetta, et al. 2021. "ERA5-Land: A State-of-the-Art Global Reanalysis Dataset for Land Applications." *Earth System Science Data* 13 (9): 4349–4383. <https://doi.org/10.5194/essd-13-4349-2021>.
- Nie, Yong, and Ainong Li. 2011. "Assessment of Alpine Wetland Dynamics from 1976–2006 in the Vicinity of Mount Everest." *Wetlands* 31 (5): 875–884. <https://doi.org/10.1007/s13157-011-0202-7>.
- Russi, Daniela. 2013. "Paper Citation: Russi D., Ten Brink P., Farmer A., Badura T., Coates D., Förster J., Kumar R. and Davidson N.(2013) *The Economics of Ecosystems and Biodiversity for Water and Wetlands*. IEEP, London and Brussels; Ramsar Secretariat, Gland."
- Shapley, Lloyd S. 1953. "A Value for N-Person Games." *Princeton University Press Princeton*: 307–317.
- Song, Yongze, Jinfeng Wang, Yong Ge, and Chengdong Xu. 2020. "An Optimal Parameters-Based Geographical Detector Model Enhances Geographic Characteristics of Explanatory Variables for Spatial Heterogeneity Analysis: Cases with Different Types of Spatial Data." *GIScience & Remote Sensing* 57 (5): 593–610. <https://doi.org/10.1080/15481603.2020.1760434>.
- Song, Yongze, and Peng Wu. 2021. "An Interactive Detector for Spatial Associations." *International Journal of Geographical Information Science* 35 (8): 1676–1701. <https://doi.org/10.1080/13658816.2021.1882680>.
- Tian, Aohua, Tingting Xu, Jay Gao, Chang Liu, and Letao Han. 2023. "Multi-Scale Spatiotemporal Wetland Loss and Its Critical Influencing Factors in China Determined Using Innovative Grid-Based GWR." *Ecological Indicators* 149 (May): 110144.
- Truong, Charles, Laurent Oudre, and Nicolas Vayatis. 2020. "Selective Review of Offline Change Point Detection Methods." *Signal Processing* 167 (February): 107299.
- Wang, Rui, Min He, and Zhenguo Niu. 2020. "Responses of Alpine Wetlands to Climate Changes on the Qinghai-Tibetan Plateau Based on Remote Sensing." *Chinese Geographical Science* 30 (2): 189–201. <https://doi.org/10.1007/s11769-020-1107-2>.
- Wang, Chao, Le Ma, Yan Zhang, Nengcheng Chen, and Wei Wang. 2022. "Spatiotemporal Dynamics of Wetlands and Their Driving Factors Based on PLS-SEM: A Case Study in Wuhan." *Science of The Total Environment* 806 (February): 151310.
- Wang, Jinfeng, and Chengdong Xu. 2017. "Geodetector: Principle and Prospective." *Acta Geographica Sinica* 72 (1): 116–134.
- Were, David, Frank Kansime, Tadesse Fetahi, Ashley Cooper, and Charles Jjuuko. 2019. "Carbon Sequestration by Wetlands: A Critical Review of Enhancement Measures for Climate Change Mitigation." *Earth Systems and Environment* 3 (2): 327–340. <https://doi.org/10.1007/s41748-019-00094-0>.
- Wu, Jianguo. 2004. "Effects of Changing Scale on Landscape Pattern Analysis: Scaling Relations." *Landscape Ecology* 19 (2): 125–138. <https://doi.org/10.1023/B:LAND.0000021711.40074.ae>.
- Xi, Yi, Shushi Peng, Philippe Ciais, and Youhua Chen. 2020. "Future Impacts of Climate Change on Inland Ramsar Wetlands." *Nature Climate Change* 11 (1): 45–51. <https://doi.org/10.1038/s41558-020-00942-2>.
- Xu, Ting, Baisha Weng, Denghua Yan, Kun Wang, Xiangnan Li, Wuxia Bi, Meng Li, Xiangjun Cheng, and Yinxue Liu. 2019. "Wetlands of International Importance: Status, Threats, and Future Protection." *International Journal of Environmental Research and Public Health* 16 (10): 1818. <https://doi.org/10.3390/ijerph16101818>.
- Xue, Zhenshan, Xianguo Lyu, Zhike Chen, Zhongsheng Zhang, Ming Jiang, Kun Zhang, and Yonglei Lyu. 2018. "Spatial and Temporal Changes of Wetlands on the Qinghai-Tibetan Plateau from the 1970s to 2010s." *Chinese Geographical Science* 28 (6): 935–945. <https://doi.org/10.1007/s11769-018-1003-1>.
- Yao, Tandong, Xiaodong Liu, Ninglian Wang, and Yafeng Shi. 2000. "Amplitude of Climatic Changes in Qinghai-Tibetan Plateau." *Chinese Science Bulletin* 45 (13): 1236–1243. <https://doi.org/10.1007/BF02886087>.
- Zhang, Guoqing, Tandong Yao, Hongjie Xie, Kun Yang, Liping Zhu, C. K. Shum, Tobias Bolch, Shuang Yi, Simon Allen, and Liguang Jiang. 2020. "Response of Tibetan Plateau Lakes to Climate Change: Trends, Patterns, and Mechanisms." *Earth-Science Reviews* 208: 103269. <https://doi.org/10.1016/j.earscirev.2020.103269>.
- Zhang, Wenjiang, Yonghong Yi, Kechao Song, John S. Kimball, and Qifeng Lu. 2016a. "Hydrological Response of Alpine Wetlands to Climate Warming in the Eastern Tibetan Plateau." *Remote Sensing* 8 (4): 336. <https://doi.org/10.3390/rs8040336>.

- Zhang, Zhen, Niklaus E. Zimmermann, Jed O. Kaplan, and Benjamin Poulter. 2016b. "Modeling Spatiotemporal Dynamics of Global Wetlands: Comprehensive Evaluation of a New Sub-Grid TOPMODEL Parameterization and Uncertainties." *Biogeosciences (online)* 13 (5): 1387–1408. <https://doi.org/10.5194/bg-13-1387-2016>.
- Zhao, Zhilong, Yili Zhang, Linshan Liu, Fenggui Liu, and Haifeng Zhang. 2015. "Recent Changes in Wetlands on the Tibetan Plateau: A Review." *Journal of Geographical Sciences* 25 (7): 879–896. <https://doi.org/10.1007/s11442-015-1208-5>.
- Zhu, Yanxin, and Yanfang Sang. 2018. "Spatial variability in the seasonal distribution of precipitation on the Tibetan Plateau." *Progress in Geography* 37 (11): 1533–1544. <https://doi.org/10.18306/dlkxjz.2018.11.009>.

A5. Measuring univariate effects in the interaction of geographical patterns

Reference: Luo,P., et al. (2023). Measuring univariate effects in the interaction of geographical patterns. International Journal of Geographical Information Science. (Under Review)

Measuring univariate effects in the interaction of geographical patterns

Peng Luo^a, Yang Li^{b,*}, Yongze Song^{c,**}, Ziqi Li^d, Liqiu Meng^a

^a*Chair of Cartography and Visual Analytics, Technical University of Munich, Munich, Germany*

^b*Key Laboratory of Remote Sensing Science, Faculty of Geographical Science, Beijing Normal University, Beijing, China*

^c*School of Design and the Built Environment, Curtin University, Perth, Australia*

^d*Department of Geography, Florida State University, USA*

Abstract

Understanding the relationships between geographical variables is a fundamental task in geographical analysis. Based on spatial dependence or heterogeneity, existing spatial methods underperform in contexts involving nonlinear relationships and intricate interactions among geographical variables. Identifying the relationships between individual variables (i.e., univariate effect) within multiple interacting variables remains challenging and long-lasting. In this study, we develop a novel model, the Geographical Pattern Interaction (GPI), based on a core premise that the spatial pattern in a response variable results from the interaction of spatial patterns in explanatory variables. The GPI model uses decision trees and Shapley value explanations to measure global and local univariate effects underlying geographical data patterns. In addition to the technical details of GPI, an empirical model of homelessness risk in Australia is used to demonstrate its utility. Results show that GPI can effectively measure nonlinear spatial associations and interactive effects that determine homelessness risk, which would be missed using existing spatial methods. We also conduct a sensitivity analysis demonstrating GPI's robustness and reliability. GPI is an interpretable and transferable tool for exploring complex spatial associations in geographical data by recognizing the often neglected spatial pattern similarities and interactions between geographic variables.

Keywords: Spatial association, spatial heterogeneity, nonlinear interaction,

*Corresponding author: Yang Li. isliyang@mail.bnu.edu.cn

**Corresponding author: Yongze Song. yongze.song@curtin.edu.au

Email addresses: peng.luo@tum.de (Peng Luo), Ziqi.Li@fsu.edu (Ziqi Li), liqiu.meng@tum.de (Liqiu Meng)

1. Introduction

One of the primary tasks in geographical analysis is to determine the relationships between geographical variables (Anselin, 1988; Brunsdon et al., 1996). Analyzing these relationships can help understand underlying data generating processes, predict future scenarios, and inform decision making. Quantitative relationships can be measured through statistical models applied to geographical data (Fotheringham et al., 2000; LeSage and Fischer, 2008). Traditional statistical models are often used to identify non-spatial relationships. Parameter estimation by using geographical samples collected from different locations, and assuming a constant relationship among geographical variables across space, however, may, result in biased results and violated assumptions. Spatial statistical approaches have been developed to explicitly address possible spatial effects present in the data. Spatial effects are normally classified into spatial dependence (Anselin, 1988) and spatial heterogeneity (Fotheringham et al., 2003; Goodchild, 2004). Spatial dependence refers to the correlation that exists between neighboring locations in geographic space, which reflects the tendency of geographical phenomena to cluster or disperse in space, and is often captured through spatial regression models (Anselin, 1995; Luo et al., 2022b). Spatial heterogeneity, on the other hand, indicates that the process generating the data may vary across different locations (Getis and Ord, 1992; De Marsily et al., 2005; Fotheringham et al., 2003). This heterogeneity may manifest that geographical phenomena exhibit different characteristics or patterns at different locations (Luo and Song, 2021). Spatial heterogeneity can be modeled in either a discrete or continuous manner. Multi-level models and spatial regimes are often used to address discrete heterogeneity (Wang and Xu, 2017; Fotheringham and Li, 2023; Anselin and Amaral, 2023), while continuous heterogeneity can be modeled using spatially varying coefficients models, such as Geographically Weighted Regression (GWR)(Fotheringham et al., 2003) and Spatial Eigenvector Filtering (Griffith and Griffith, 2003).

Previous spatial models are predominantly constructed based on the two spatial effects mentioned above to reveal spatial association, but there remain some unresolved

30 issues. First, they often fail to consider the complex interactions among explana-
31 tory variables (Song et al., 2020; Zhang et al., 2023), and require that the residuals
32 from different variables are independent (Anselin, 1989). The spatial distribution of
33 geographical variables can be influenced by the interaction of multiple explanatory
34 variables. For example, the function between the response variable and an individual
35 explanatory variable X_a , as represented by $f(X_1)$ may include interactions with other
36 variables (e.g. X_b): $f(X_1 \cap X_2)$. It is crucial to comprehend how individual variables
37 relate to one another (the univariate effect) when there are several interacting variables
38 at play. Ignoring the interaction among variables could cause the univariate effect to
39 be over- or underestimated. Second, current spatial models heavily rely on linearity
40 assumptions (Comber et al., 2021; Li, 2022). The relationships between geographical
41 variables often exhibit significant non-linearity. For instance, a moderate increase in
42 temperature can enhance the growth of certain plants, but excessively high tempera-
43 tures may inhibit their growth (Zhu et al., 2021b). Third, current models often impose
44 strict statistical assumptions on data distribution (e.g., Normal, Poisson), which are
45 usually violated in real geospatial dataset (Arbia, 2006). However, geographical data is
46 often influenced by spatial dependency resulting non i.i.d.(independent and identically
47 distributed) samples. In summary, current spatial models for detecting relationships
48 commonly overlook variable interaction effects and impose inflexible statistical assump-
49 tions on functional forms (e.g., linearity) and data distributions.

50 The limitations mentioned above arise from the modeling approach to spatial ef-
51 fects. Most models for detecting spatial relationships assume that spatial effects, such
52 as spatial heterogeneity, are continuously present across space. They rely on the val-
53 ues of geographical variables at each sample point to model the relationships between
54 them. This process typically requires statistical assumptions, such as linearity and nor-
55 mal distribution (Gao et al., 2022). However, in reality, spatial heterogeneity can be
56 modeled not only continuously across geographical space but also discretely (Anselin
57 and Amaral, 2023). Concepts like spatial regimes and stratified spatial heterogeneity
58 have been introduced to describe this discrete heterogeneity (Anselin, 2010). Within
59 the theoretical framework for discretely modeling spatial heterogeneity, regions with
60 heterogeneity are subdivided into multiple homogeneous subregions (Guo et al., 2023).

61 In each subregion, model parameters remain consistent, indicating a presumed station-
62 ary spatial relationship.

63 This study models discrete spatial heterogeneity to identify spatial correlations,
64 with the expectation of not imposing overly strong statistical assumptions on spatial
65 data. We introduce a new way of thinking about spatial relationships in line with
66 people’s intuitive understanding of the world: the more similar the spatial distribution
67 patterns of two geographical variables, the stronger their relationship may be. The
68 extent to which an explanatory variable X’s spatial distribution influences the spatial
69 distribution of the response variable Y can reflect their correlation. Our aim is to
70 incorporate this assumption into spatial correlation analysis by evaluating the spatial
71 pattern similarity among geographical variables to analyze their mutual relationships
72 and interaction strength.

73 In addition to proposing relatively loose statistical assumptions, we attempt to be
74 able to reveal interactions among multiple explanatory variables. The geographically
75 optimal zones-based heterogeneity model (GOZH) was developed to address this is-
76 sue (Luo et al., 2022a). It has been proven effective in identifying spatial associations
77 by detecting the spatial patterns of geographical variables. Its purpose is to achieve
78 geographically optimal zones for every combination of variables and to determine the
79 geographically optimal partition given all explanatory variables. GOZH does not rely
80 on statistical assumptions about data distribution, such as assuming a normal distribu-
81 tion. Furthermore, GOZH has the capability to detect non-linear relationships between
82 variables.

83 However, GOZH cannot reveal global univariate effects under conditions of pattern
84 interaction. It can provide an overall assessment of the combined impact of multiple
85 explanatory variables on the response variable, but it cannot calculate the individual
86 contribution of a specific variable within this interaction. Secondly, it cannot explore
87 the local univariate effects under conditions of pattern interaction. One of the key
88 distinctions between models based on discrete heterogeneity (e.g. GOZH model) and
89 models based on continuous heterogeneity (e.g. GWR model) is that the former cannot
90 estimate local spatial non-stationary parameters. Due to this limitation, GOZH cannot
91 explore the nonlinearity of local univariate effects.

92 Based on the above discussion, there is a critical need to develop a new approach for
93 measuring univariate effects in the interaction of geographical patterns. This approach
94 should simultaneously achieve two main objectives: (i) to identify the overall correlation
95 between explanatory variables and response variables (global effect), and (ii) to reveal
96 the spatial variance of their relationship (local effect).

97 Consequently, this study develops a geographical pattern interaction (GPI) model
98 to identify the univariate effects under the condition of pattern interaction. GPI model
99 works in three steps. The first one is to generate the GPI of the response variable given
100 the spatial pattern of multiple explanatory variables. The GOZH model is used to de-
101 tect this spatial pattern. The second is to calculate the global univariate effects in GPI.
102 For each geographic partitioning under each combination of variables, we calculate the
103 variance of each partition and the overall relationship between the explanatory and
104 response variables. This analysis includes the interaction between a single explanatory
105 variable and multiple explanatory variables, using the concept of spatially stratified
106 heterogeneity. We then introduce Shapley value, an interpretable machine-learning al-
107 gorithm based on game theory, to quantify the contribution of a single variable in the
108 interaction of multiple explanatory variables ([Štrumbelj and Kononenko, 2014](#); [Lund-
109 berg et al., 2020](#); [Li, 2022](#); [Li et al., 2023](#); [Li, 2023](#)). The third is to calculate local
110 univariate effects in GPI, including local effects of GPI, local univariate effects, and
111 their characteristics of nonlinearity, local dominant variables, and bi-variate effects.
112 We computed the means of the response variables for each geographical partition of
113 the region under different combinations of variables, and then used Shapley to explore
114 the contribution of different variables to the classification of the region as a geograph-
115 ically optimal partition. This contribution represents the relationship between the
116 explanatory variables and the response variables in this region.

117 The remainder of this paper is organized as follows: Section 2 introduces the concept
118 and framework of the GPI model; Section 3 presents a case study of applying the GPI
119 to analyze the risk of homelessness in Australia; Section 4 demonstrates the results of
120 the case study. We discuss the methodological contributions of the GPI in Section 5
121 and conclude the study in Section 6.

122 2. Geographical pattern interaction (GPI)

123 2.1. The concept

124 The essence of GPI is based on the hypothesis that the interaction between geo-
125 graphical patterns can indicate spatial association. The distribution of the response
126 variable is determined by a series interactions among multiple explanatory variables.
127 For a specific explanatory variable X , its impact on the spatial pattern of the re-
128 sponse variable Y represents the spatial association between X and Y . Spatial pattern
129 can be characterized in various ways. Normally, the distribution of a geographical
130 variable exhibits a spatial pattern: its values can be grouped into several relatively ho-
131 mogeneous subregions, where values are similar within each subregion and dissimilar
132 between different subregions (Anselin and Amaral, 2023). The proper description of
133 the spatial pattern of each geographic variable can facilitate the exploration of their
134 spatial association. We find that the stronger the spatial pattern interaction driven by
135 explanatory variables, i.e. higher similarity within subregions and greater dissimilarity
136 between subregions in the spatial pattern of the response variable, the stronger the
137 spatial association between these explanatory variables and the response variable.

138 For instance, we are interested in understanding the association between elevation
139 (explanatory variable) and temperature (response variable). We divide the area based
140 on elevation into high-elevation and low-elevation regions. When we observe that
141 high-elevation regions generally have lower temperatures, and low-elevation regions
142 generally have higher temperatures, it is reasonable to infer that temperature is strongly
143 influenced by (or associated with) elevation.

144 More often, a geographical phenomenon of interest is associated with multiple ex-
145 planatory variables and their spatial associations can be determined in a similar manner
146 as shown in Figure 1. Suppose the response variable Y is determined by two explana-
147 tory variables, X_a and X_b . The value of the response variable, Y , is indicated by the
148 color value, while explanatory variables X_a and X_b are distinguished by different color
149 hues to represent different distributions or categories of values. For example, three
150 regions of X_a —red, blue, and green—may represent three different states or levels of
151 the variable. We can determine the geographic interaction patterns of Y based on X_a ,
152 X_b , and the interaction between X_a and X_b . The distribution of X_a allows for the

153 spatial division of Y into three regions, with mean values of 7.5, 15, and 10, respec-
154 tively. We utilize the index q to assess the spatial stratification heterogeneity of this
155 case, which measures the extent to which the spatial distribution of X can explain (i.e.
156 distinguish) the distribution of Y . The q value for X_a is 0.661 with a p -value of less
157 than 0.01, indicating a significant spatial association between X_a and Y . On the other
158 hand, based on X_b 's distribution, Y is divided into two subregions with a mean value
159 of 10, indicating that X_b alone is ineffective in distinctly partitioning Y . The q -value
160 for X_b is 0.002 with a p -value of 0.783, suggesting no significant spatial association
161 with Y . However, when X_a and X_b interact, the geographic pattern of Y is divided
162 into three subregions with mean values of 5, 15, and 10, achieving a perfect separation.
163 The q -value is 0.925 with a p -value of less than 0.01, implying a very strong spatial
164 association between Y and the combination of X_a and X_b . This demonstrates that the
165 collaborative interaction of X_a and X_b can explain 92.5% of the spatial variability in
166 Y . It also indicates that some explanatory variables, like X_b , cannot independently
167 affect the response variable. However, when they interact with other variables, their
168 explanatory power is significantly enhanced. This phenomenon is common in the field
169 of geography but has been difficult to reveal with previous models.

170 After detecting the interaction effect of multiple explanatory variables, we aim to
171 identify the univariate effects in geographical pattern interaction, which represent the
172 contribution of a specific explanatory variable in the interaction. In the proposed GPI
173 model, we introduce the idea of game theory. To evaluate the contribution of a specific
174 explanatory variable to the response variable, we assess how much it helps achieve
175 the spatial pattern mentioned earlier: different combinations of explanatory variables
176 result in various grouping methods. For a particular explanatory variable, such as X ,
177 the decision to include or exclude it may make difference. We argue that the changes in
178 between-group variance among different groupings can reflect the overall contribution
179 (global) of the explanatory variable X to the response variable. Moreover, fluctuations
180 in the means assigned to each location or region's group can provide insight into the
181 contribution (local) of X to the response variable at that specific location.

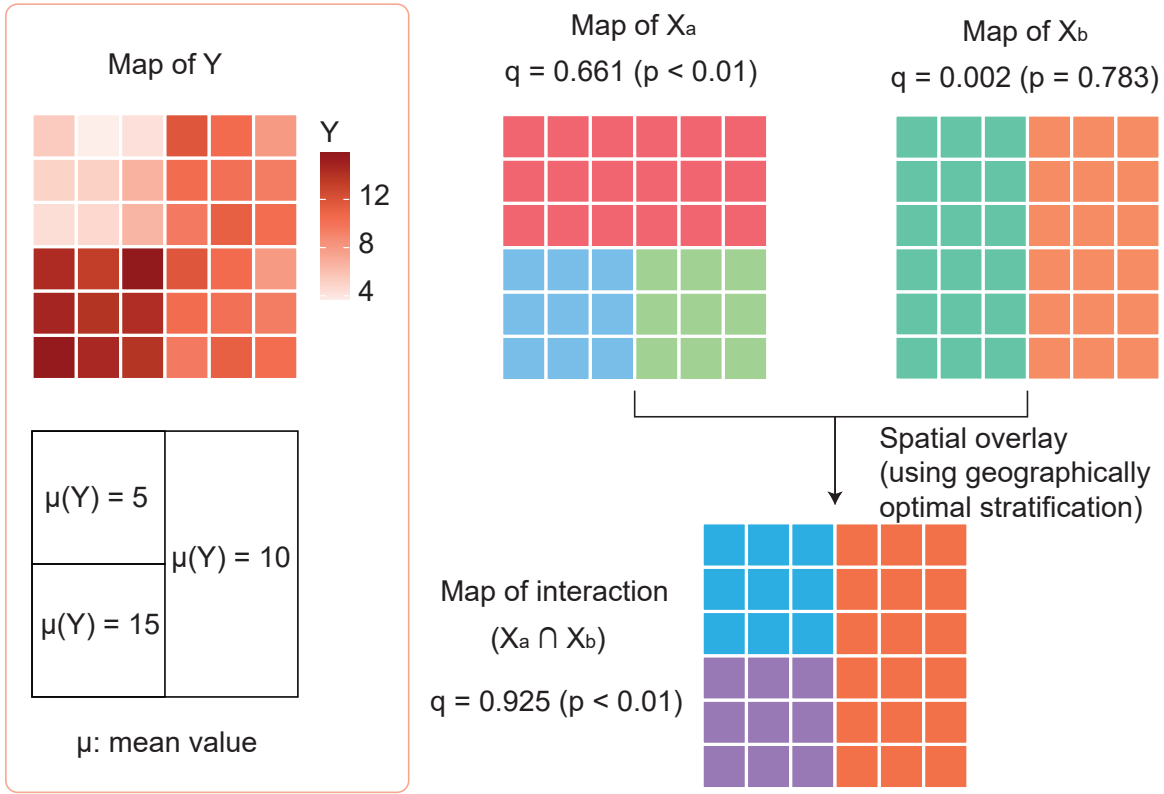


Figure 1: An example of identifying the spatial association according to the pattern interaction. The black lines represent the division of the response variable space based on explanatory variables.

182 *2.2. The GPI model*

183 Figure 2 shows the framework of the GPI model designed to explore the interac-
 184 tion of distribution patterns among geographical variables, including both global and
 185 local spatial relationships among various variables, and to assess the impact of each
 186 individual variable within this interaction. In the framework, a response variable and
 187 three explanatory variables are annotated as Y and X_a , X_b , and X_c .

188 In the first step, the model generates the geographical pattern interaction of the
 189 response variable. The spatial distribution of Y is divided into several homogeneous
 190 subzones given the interaction of the explanatory variables. The model is able to min-
 191 imize the sum of the within-zone variance of the response variable across all subzones.
 192 In the second step, the model calculates the global univariate effects in the GPI. It
 193 reveals the overall impact of each explanatory variable on the response variable. It also
 194 detects the contribution of each explanatory variable in the GPI.

195 In the third step, the model detects the local univariate effects in GPI. Firstly, the

196 spatial distribution of local effects of GPI is mapped. Secondly, at each spatial position,
 197 the model discerns the spatial correlation between each individual variable X and the
 198 response variable Y . In this manner, the spatial distribution of these correlations can
 199 be visualized. Thirdly, the model explores the nonlinearity of the contribution of
 200 each explanatory variable. Fourthly, the locally dominant variable is detected which
 201 represent the explanatory variable with the strongest association with the response
 202 variable. Finally, at each specific location, the model delves into the interactive effects
 203 of different variables.

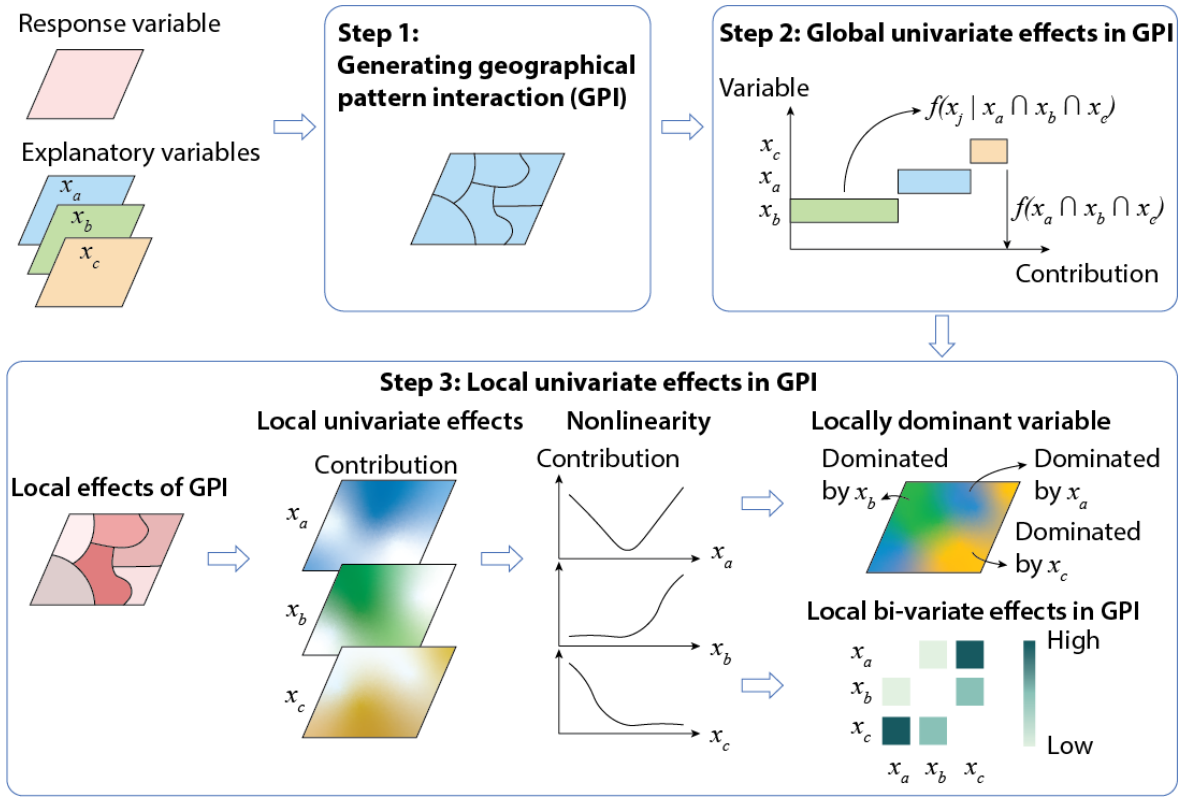


Figure 2: Flowchart of geographical pattern interaction (GPI) model

204 *2.2.1. GPI generation*

205 To clarify the spatial pattern, the first step is to map the interaction of geographical
 206 patterns with the response variables. We define the outcomes of these geographical
 207 pattern interactions as homogeneous regions derived from spatial distribution. Within
 208 each region, the homogeneity should ideally approach optimality (Luo et al., 2021).
 209 The model should minimize the total variance among these regions, as is measured by
 210 the PD value:

$$PD = \min (SSW_{X,D}) = \min \left\{ \sum_{z=1}^h \sum_{j=1}^{N_z} (y_{z,j} - \bar{c}_z)^2 \right\} \quad (1)$$

211 where X is one or multiple explanatory variables, D is the stratified variable for de-
 212 scribing geographical strata, and $SSW_{X,D}$ is the sum of squares within geographical
 213 strata that are recorded as D and determined by explanatory variable X . $y_{z,j}$ and \bar{c}_z are
 214 the j th observation and mean values of response variables in stratum z , respectively.

215 The given Equation 1 presents a challenging problem in the context of nondetermin-
 216 istic polynomial-time complete (NP-complete) problem, making it difficult to ascertain
 217 a global optimum solution. To address this equation, we employ a heuristic approach
 218 involving spatial explanatory variables to perform a gradual spatial discretization. The
 219 classification and regression tree (CART) model is introduced to conduct the step-wise
 220 process:

$$\min_{k,s} \left\{ \sum_{x_i \in Z_1(k,s)} (y_i - \bar{d}_1)^2 + \sum_{x_i \in Z_2(k,s)} (y_i - \bar{d}_2)^2 \right\} \quad (2)$$

221 where \bar{d}_1 and \bar{d}_2 are the average values of response variable in subzone Z_1 and Z_2 ,
 222 respectively.

223 In GPI, the choice of spatial discretization method depends on various research
 224 questions and data. In this study, we use CART because a decision tree-based model
 225 can offer good interpretability. More details regarding to the spatial discretization
 226 process can refer to GOZH model (Luo et al., 2022a).

227 2.2.2. Global univariate effects in GPI

228 Global geographical association (for whole study area \mathbf{u}) between explanatory vari-
 229 able X_j and Y is described as :

$$G = f(X_j(\mathbf{u}) | X_1 \cap X_2 \cap \dots \cap X_n) = \varphi_{X_j}(PD) \quad (3)$$

230 where $\mathbf{u} = [u]$, indicates the whole study area containing all location u , $f(x)$ is the
 231 impact of geo-interaction, φ is the function to calculate the contribution of a single
 232 explanatory variable. In this study, we specifically apply the Shapley values to quantify
 233 the univariate contribution to the response variable. Shapley values is a method to

234 fairly distribute the “payout” among players in game theory (Shapley et al., 1953). It
 235 can be calculated as:

$$\varphi_{X_j}(PD) = \left| \sum_{s \in C \setminus \{X_j\}} \frac{|S|!(|C| - |S| - 1)!}{|C|!} (PD(S \cup \{X_j\}) - PD(S)) \right| \quad (4)$$

236 where, C is a set of all variables excluding X_j , S is a subset of all possible combinations
 237 of C . $|S|$ and $|C|$ represent the number of variables in the set.

238 Specifically, we iterate through all possible variable combination (M combinations
 239 in total, $M = 2^n - n - 1$) to compute the PD value based on formula 1 and 2. Then,
 240 the univariate effects are calculated as in formula 4. The $\varphi_{X_j}(PD)$ is the weighted
 241 average of the gain in PD values attributable to variable X_j under all combinations.
 242 In the end, Shapley value of the PD of each variable is calculated to quantify the effect
 243 of this variable under GPI.

244 2.2.3. Local univariate effects under GPI

245 Local univariate effects in GPI contain five components: (i) overall local effects
 246 (average value of each subzone); (ii) local univariate effects under the condition of
 247 GPI; (iii) nonlinearity of local univariate effects; (iv) identifying predominant local
 248 variables; and (v) local interaction effects under the condition of GPI.

249 The first component is the overall local effects. To a specific location u , the local
 250 effect of a combination of explanatory variable $C(x_1, x_2 \dots x_j \mid 1 \leq j \leq n)$ is calculated
 251 as:

$$l_u(C) = \frac{\text{Sum}_z(C)}{n_z} \quad (5)$$

252 where z is a stratum of the study area.

253 The second component is employing Shapley values to calculate local univariate
 254 effects under the condition of GPI, which is described as :

$$L = f(X_j(u) | X_1 \cap X_2 \cap \dots \cap X_n) = \varphi(l_u) \quad (6)$$

255 where $f(x)$ is the impact of geo-interaction, φ is the function to calculate the
 256 contribution of a single explanatory variable. We also apply the Shapley values to
 257 quantify the local univariate effects. The difference between the calculation of the

258 global univariate effect is that, in this case, the input is an indicator used to describe
 259 the local spatial pattern in location u (as specified in formula 5). Therefore, the
 260 univariate effects for each location are calculated as follows:

$$\varphi_{x_j}(l_u) = \sum_{s \in C \setminus \{x_j\}} \frac{|S|!(|C| - |S| - 1)!}{|C|!} (l_u(S \cup \{x_j\}) - l_u(S)) \quad (7)$$

261 where u is a location.

262 Thirdly, we explore the non-linear relationships between variables. We group the ex-
 263 planatory variables using quantiles and then examined the corresponding local Shapley
 264 values. This allows us to obtain non-linear impact curves of the explanatory variables
 265 on the dependent variable. Fourthly, we calculate the spatial determinants of each
 266 location, which is the variable most strongly correlated with the response variable.

267 Finally, we compute local interaction effects as the Shapley interaction values do.
 268 Shapley interaction values further decompose local effect into main effect and interac-
 269 tion effect. The main effect refers to the individual contribution of each variable, which
 270 is independent of other variables and can help us understand the importance of each
 271 variable and its impact on the explained variable. The interaction effect considers the
 272 synergy between variables. It tells us what kind of impact they have on the explained
 273 variable when multiple variables are considered together. In other words, the interac-
 274 tion effect refers to the additional contribution from the interaction of the variables.
 275 The value is calculated as follows:

$$\varphi(x_i \cap x_j) = \left| \sum_{s \in M \setminus \{x_i, x_j\}} \frac{|S|!(|M| - |S| - 2)!}{(|M| - 1)!} (\Delta l_u(S, x_i, x_j)) \right| \quad (8)$$

where, $\varphi(x_i \cap x_j)$ represents the interaction value of the coalition of variables x_i and
 x_j , $\Delta l_u(S, x_i, x_j)$ represents the additional reward obtained from the coalition:

$$\Delta l_u(S, x_i, x_j) = l_u(S, x_i, x_j) - l_u(S, x_j) - l_u(S, x_i) + l_u(S) \quad (9)$$

276 3. Case study: a GPI to identify factors influencing the risk of homelessness

277 3.1. Study area and data

278 We implemented our model to explore the spatial associations in homelessness risk
 279 data from Australia in 2016 (Parkinson et al., 2019). The data were from Australian

280 Bureau of Statistics, recording the number of people at risk of homelessness per 10,000
 281 at the Statistical Area Level 3 (SA3) level (Australian Bureau of Statistics, 2016). The
 282 risk indicator of homelessness represents the percentage of the population lacking a
 283 permanent residence, compelled to dwell in open spaces, temporary shelters, or inade-
 284 quate living conditions. This metric is widely employed to evaluate the socio-economic
 285 circumstances and fairness within a country or region. Individuals who lose stable
 286 housing often encounter challenges such as diminished employment prospects and in-
 287 creased health problems, ultimately resulting in increased social security expenditures
 288 and economic burdens. Comprehending and addressing the risk of homelessness is
 289 an essential step toward a just and compassionate society in Australia. By reducing
 290 this risk, we can foster greater social integration, thereby promoting social harmony
 291 (Caton et al., 2005). A precise understanding of the extent and causes of homelessness
 292 empowers the government to formulate more effective social policies and intervention
 293 strategies to help those most in need.

294 Figure 3 shows the spatial distribution of homelessness risk in Australia. The
 295 homeless risk is high in the central and northern regions, whereas low in the eastern
 296 and western coastal regions. It has an evident spatial heterogeneity inside several major
 297 cities, which is high in the city centers and nearby suburbs.

298 For the analysis of the spatial association, data of seven socio-economic indicators
 299 were collected as the explanatory variables that are potentially related to the homeless-
 300 ness risk. They are population size, population density, unemployment rate, proportion
 301 of dwellings without internet connection, rental costs, mortgage rate, and commuting
 302 distance to work. All variables were collected at the SA3 level.

Table 1: Explanatory variables of the homelessness risk

Name	Code	Description
population	pop	estimated resident population
population density	popdens	population density (persons/ km^2)
unemployment rate	unemployment	unemployment rate (%)
non internet rate	noninternet	proportion of dwelling without Internet access (%)
rental payment	rentalpay	median weekly household rental payment (\$)
mortgage affordability	affordmort	Households where mortgage repayments are more than 30% of imputed household income
distance to work	diswork	average commuting distance from place of usual residence to work (km)

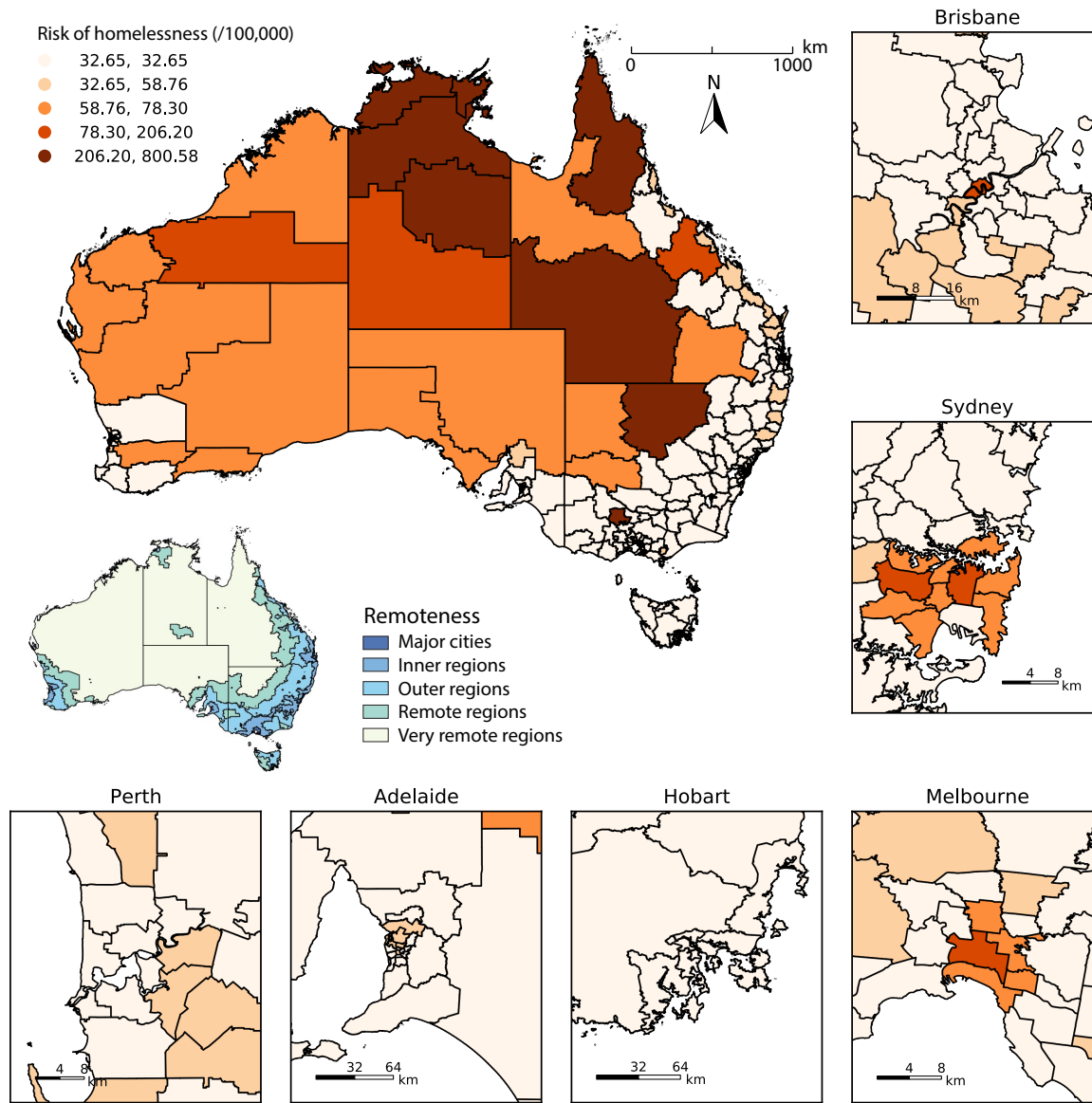


Figure 3: The spatial distribution of homeless risk in Australia. The small map displays the remoteness classification of Australia.

303 *3.2. The workflow of applying GPI for the case study*

304 The proposed GPI model was employed to investigate the factors associated with
 305 the homelessness risk in Australia. First, we conduct data preprocessing and generate
 306 the geographical pattern interaction. The distribution of the homelessness risk in
 307 Australia was divided into several subzones based on the interaction among explanatory
 308 variables. This geographically optimized zoning forms the basis for our subsequent
 309 analysis using the GPI model. Second, we assessed the global univariate effects in
 310 GPI, calculating the spatial association of individual variable with the homelessness

311 risk. Third, we evaluated the local univariate effects in GPI. The spatial locally impact
 312 of each individual variable on homelessness risk was estimated. Finally, to validate the
 313 model's performance and robustness, a sensitivity analysis was conducted using the
 314 leave-one-out method. In this approach, we systematically removed one region at a
 315 time from the model's input data and examined the resultant changes in dominant
 316 variables. The model sensitivity was gauged by measuring the percentage change in
 317 these dominant variables across each region.

318 4. Results

319 4.1. The GPI of homelessness-related variables

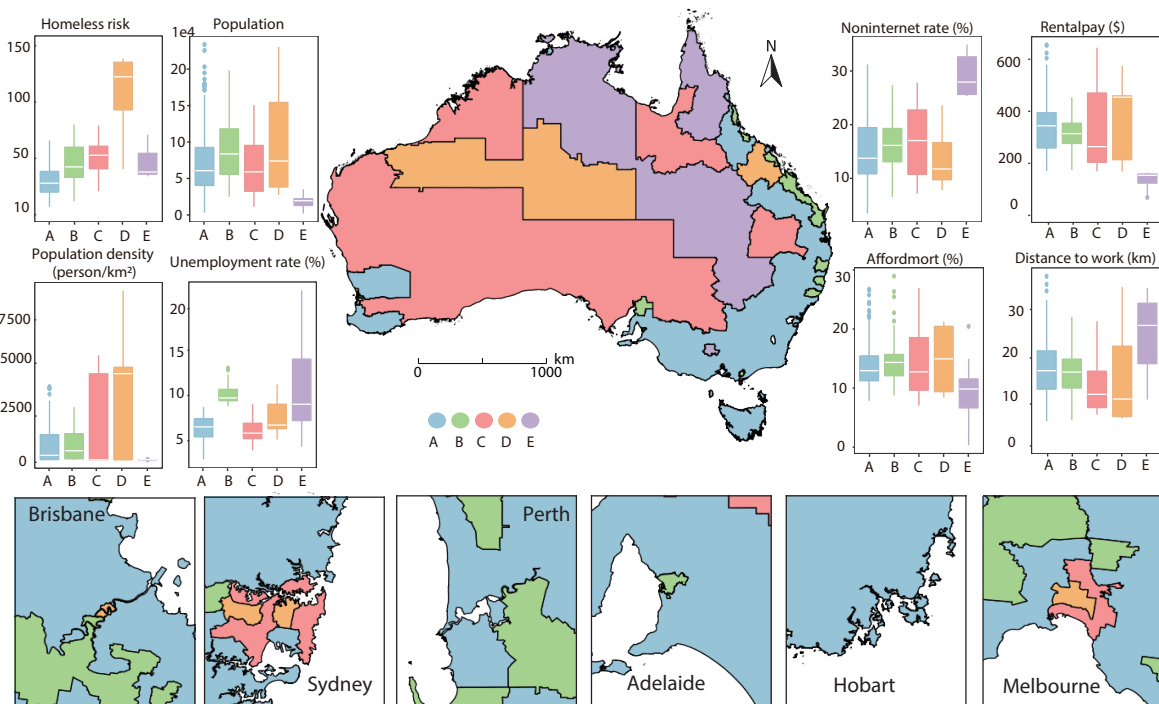


Figure 4: The geographical pattern interaction: geographically optimal zones of homelessness risk

320 Figure 4 depicts the optimal geographic zones of homeless risk in Australia, de-
 321 termined by seven explanatory variables. Five distinct sub-zones with homogeneous
 322 homelessness risk were identified. Zone A, having the lowest homelessness risk with
 323 an average value of 32.7, primarily comprises eastern cities. Zone B, mainly located
 324 in the eastern coastal regions, has a slightly higher average homelessness risk of 58.8
 325 than Zone A and the highest average unemployment rate among the five zones at

326 9.8%. Zone C encompasses most of the vast western inland areas, and also includes
 327 some regions in major cities like Sydney and Melbourne. Zone D, with the most severe
 328 homelessness problem (average risk of 206.2), includes several inner areas and inner
 329 cities of Sydney and Melbourne. Zone E, characterized by a sparse population (aver-
 330 age population density of 0.9), has the highest average non-internet rate (29.1%) and
 331 distance to work (24.2km), as well as the lowest average rental payment (128.9\$) and
 332 mortgage affordability (9.1%).

333 *4.2. Global univariate effects in GPI*

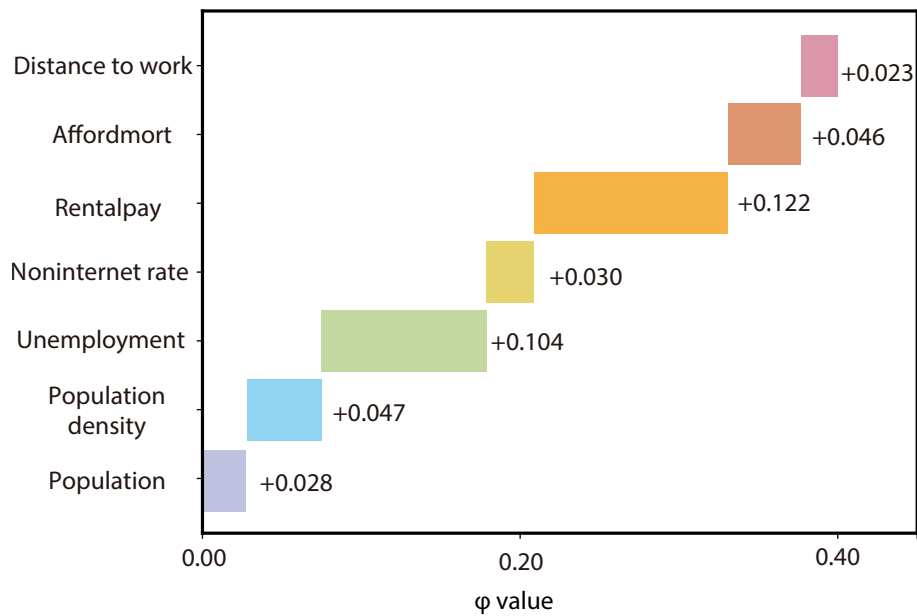


Figure 5: The global univariate effects in GPI

334 Figure 5 illustrates the strength of the correlations between individual variables
 335 and the homelessness risk. The interaction of seven explanatory variables explain to-
 336 tally 40% of the spatial disparities of homelessness risk. Among them, rental payment
 337 has the strongest association, with a φ value of 0.122. Next is unemployment rate,
 338 which can explain 10.4% of the homelessness risk. High rental costs may reduce the af-
 339 fordability of some individuals to housing expenses, forcing them to become homeless.
 340 Therefore, rental payment is likely a contributing factor to the increase of the home-
 341 lessness risk. Similarly, as the unemployment rate rises, more people may lose their
 342 jobs, struggle to sustain their livelihoods, and face the risk of homelessness. Hence, an
 343 increase in the unemployment rate might lead to a rise in the homelessness risk.

344 The associations between population density, mortgage affordability, the rate of no
345 internet, and the homelessness risk are similar, with φ values of 0.047, 0.046, and 0.030,
346 respectively. Poorer mortgage affordability could be associated with a higher risk. If
347 some individuals are unable to pay their mortgages, they may lose homeownership.
348 Higher population density refers to a greater number of people residing in a given
349 area. Areas with higher population density may be more susceptible to experiencing
350 a higher homelessness risk. This could be due to increased competition for limited
351 housing resources and higher rental costs, making it difficult for some individuals to
352 afford housing. The lack of internet access could be related to the homelessness risk
353 by preventing individuals from accessing job information, educational resources, and
354 social assistance, thereby increasing the homelessness risk.

355 *4.3. Local univariate effects in GPI*

356 *4.3.1. Local effects of GPI*

357 Figure 6 shows the local effects of GPI among all explanatory variables. The spatial
358 distribution of explanatory variables exhibit a significant clustering pattern in their
359 local effect on homelessness risk. The Moran's I is 0.35, with a Z score of 9.78. The local
360 effect in most southeast coastal regions is relatively low (32.65), while its considerably
361 high (206.20) in the city centers of Sydney, Melbourne, and Brisbane. This reveals a
362 strong association between explanatory variables and homelessness risk in the major
363 urban areas.

364 *4.3.2. Local univariate contributions*

365 Figure 7 shows the local univariate contribution of each explanatory variables. In
366 coastal and urban areas (major cities, inner, outer regions), population may be a
367 primary factor contributing to homelessness risk since population influences housing
368 demand. In addition, the mortgage affordability plays a crucial role because the high
369 cost of homeownership can exert financial pressure on low-income households, thereby
370 increasing homelessness risk. In inland and remote areas (remote, very remote), popu-
371 lation density and the proportion of rental payments are more closely associated with
372 the homeless risk. Population density can be related to the balance between housing
373 supply and demand, and in areas with lower population density, it may be challenging

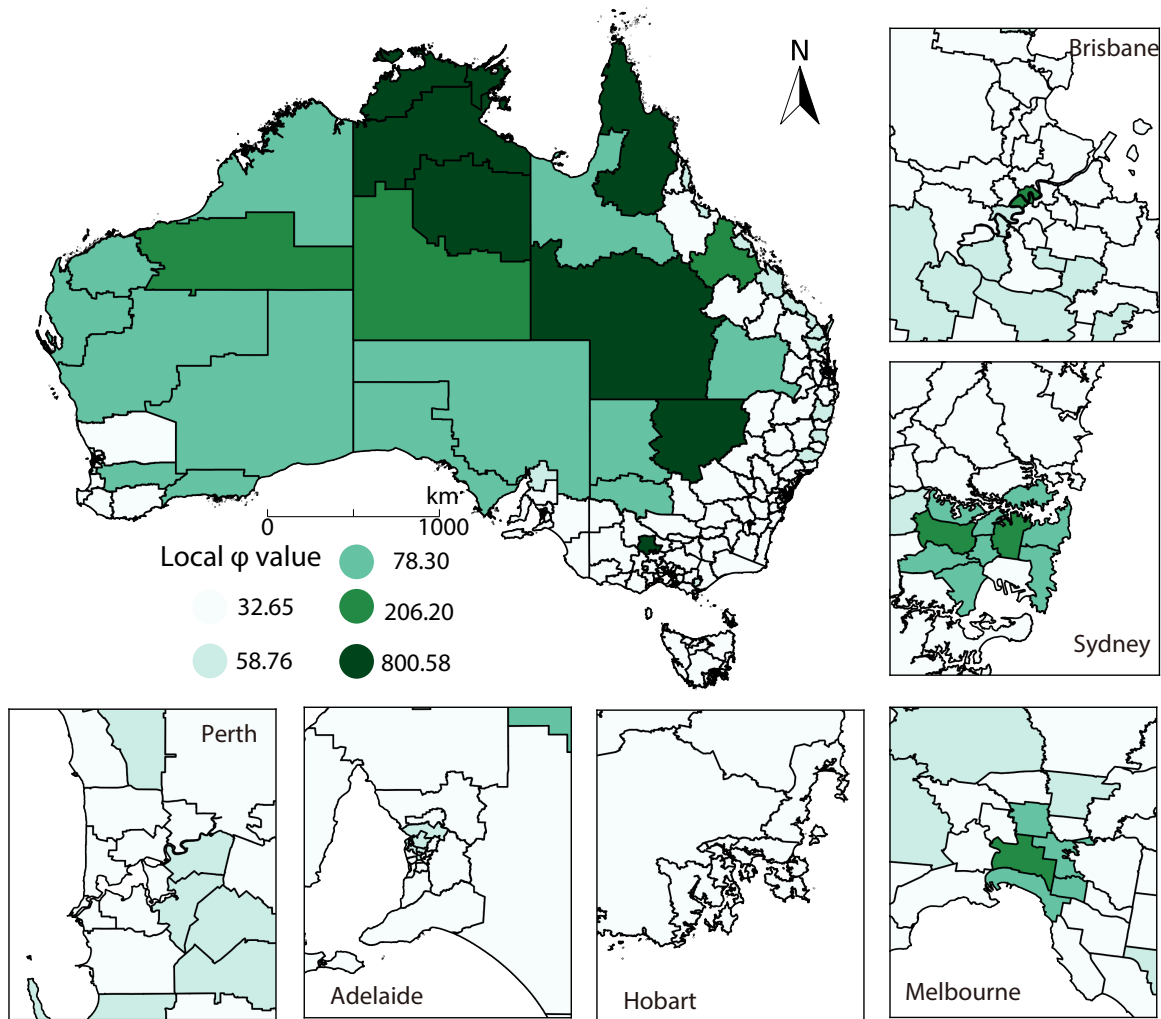


Figure 6: Local effects of GPI

374 to provide an adequate range of housing choices. In these regions, the burden of high
 375 rental payments can significantly affect the economic stability of low-income house-
 376 holds. Due to limited opportunities for homeownership, many individuals may rely on
 377 renting, and high rental costs can lead to financial instability, thereby increasing the
 378 risk of homelessness.

379 4.3.3. Non-linear effects

380 Figure 8 indicate the non-linear relationship between explanatory variables and the
 381 homelessness risk. Figure 8 (a) reveals relationship between the population and the
 382 homelessness risk. For areas with a smaller population, population size significantly
 383 influences the homelessness risk. However, as the population increases, this correlation
 384 diminishes. Below 50,000 in population, a strong association exists, but above this

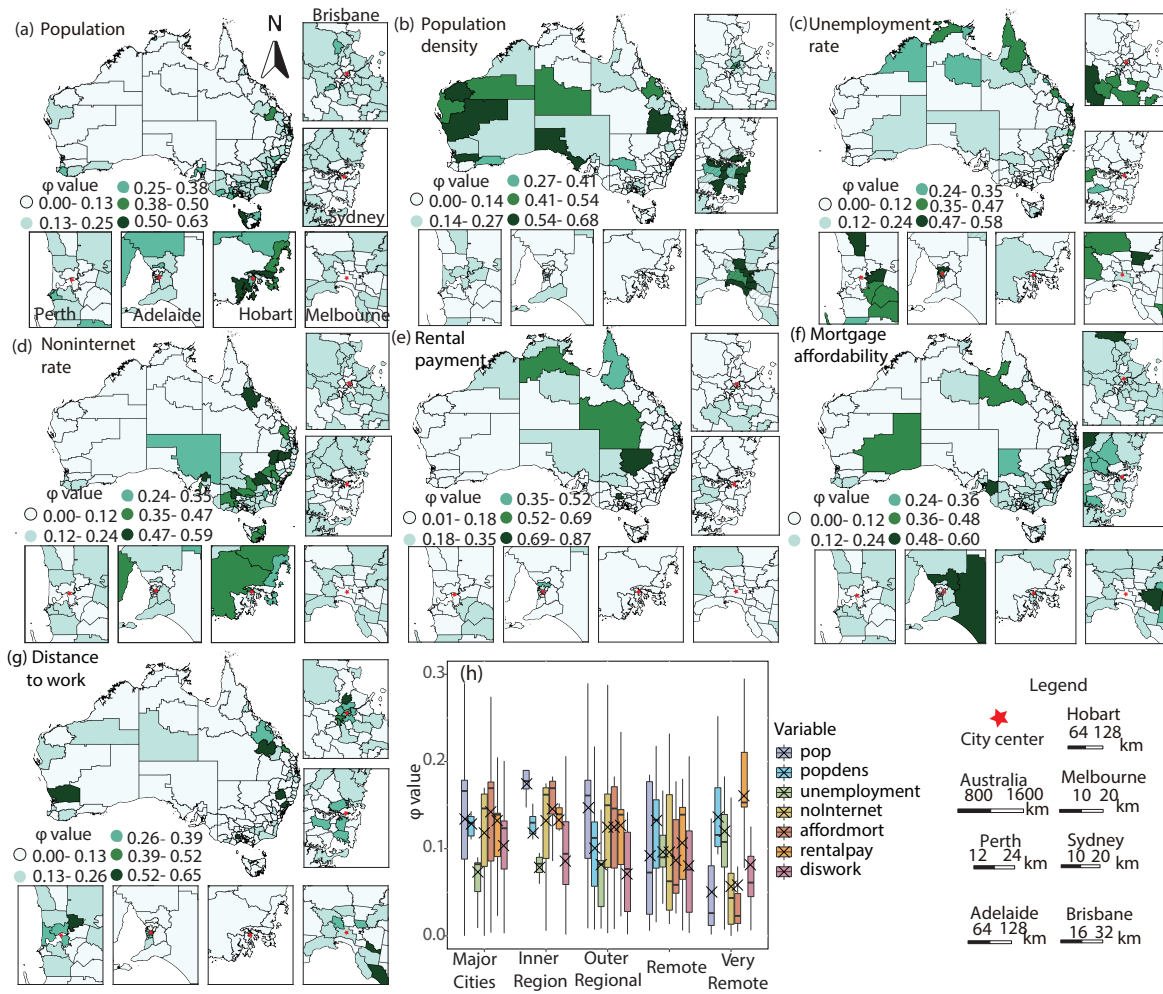


Figure 7: The local univariate effects in GPI: (a) Population; (b) Population density; (c) Unemployment rate; (d) Non-internet rate; (e) Rental payment; (f) Mortgage affordability; (g) Distance to work

385 threshold, the correlation stabilizes at a lower level. This suggests that beyond a
 386 population of around 50,000, community complexity and diversity increase, leading to
 387 the saturation of the population's impact on the homelessness risk. In Figure 8 (b), we
 388 observed that higher population density in regions has a more pronounced influence on
 389 the homelessness risk. When population density falls below 2,500 people per square
 390 kilometer, the influence remains relatively stable. In sparsely populated areas, the
 391 homelessness risk appears less sensitive to population density. In summary, smaller
 392 communities with limited social resources, including housing and social services, are
 393 more sensitive to changes in population size, resulting in noticeable effects on the
 394 homelessness risk. Conversely, in high-density areas, where social resources are more
 395 abundant, social issues and competition may lead to a more pronounced impact of

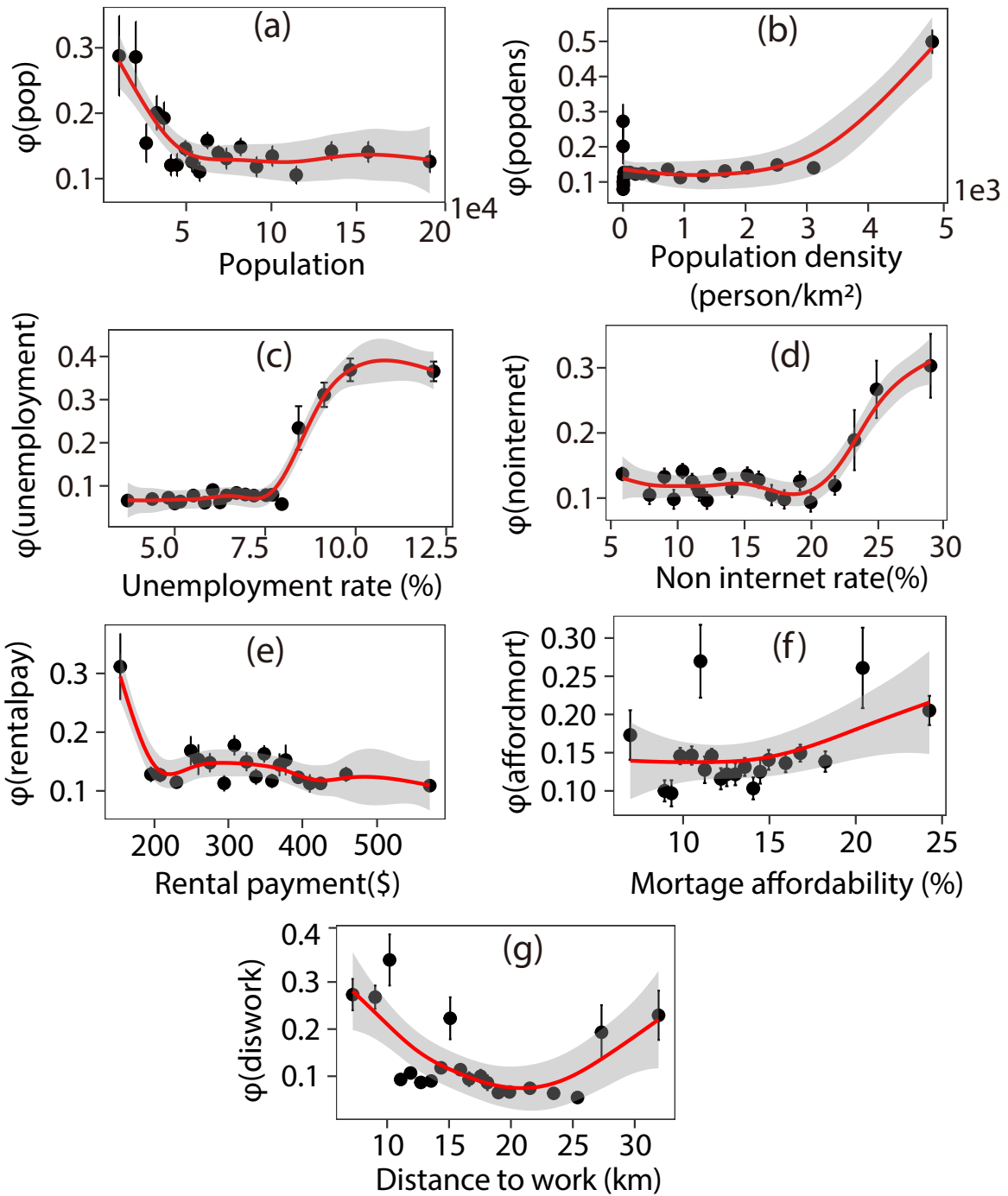


Figure 8: The non-linearity of univariate effects in GPI caused by: (a) population; (b) population density; (c)unemployment rate; (d)noninternet; (e)rental payment; (f) mortgage affordability; (g) diswork

396 population density on the homeless risk.

397 The unemployment rate (Figure 8 (c)) and the rate of no internet (Figure 7 (d))
 398 have similar effects on homelessness. When the unemployment rate is low (less than

399 7.5%) or the rate of no internet is low (less than 20%), their impact on homelessness
400 is weak. However, as the unemployment rate or the no internet rate rises, they sig-
401 nificantly increase homelessness, accounting for around 40% and 30% of homelessness
402 risk, respectively. These relationships are influenced by economic conditions, social
403 safety nets, and the housing market. Low unemployment rate in a thriving economy
404 reduces homelessness, while rising unemployment rate leads to job insecurity and hous-
405 ing mortgage affordability issues, increasing homelessness. Moreover, a rising rate of no
406 internet can impede individuals in finding job opportunities, as many job searches oc-
407 cur online, further increasing homelessness risk for those without internet access. The
408 abrupt changes at critical thresholds (7.5% unemployment rate and 20% no internet
409 rate) are due to the non-linear nature of the economic system. Below these thresholds,
410 the system is stable with milder interactions, reducing the impact of homelessness. Sur-
411 passing these thresholds leads to sudden instability, driven by complex factors, reduced
412 market confidence, and economic downturn, leading to more homelessness. Economic
413 feedback loops worsen employment prospects, creating a self-reinforcing cycle.

414 Figure 8 (e) illustrates the relationship between rental payments and homelessness
415 risk. In regions characterized by low rental payments, the impact of rental payments
416 is substantial. Furthermore, this impact diminishes as rental payments increase. How-
417 ever, in regions with higher rental payments (around 200 units), the correlation between
418 rental payments and homelessness risk is weaker, indicating that an increase in rental
419 payments does not necessarily lead to a higher homelessness rate. Our findings high-
420 light the vulnerability of low-income households to homelessness as a consequence of
421 rent increases. This underscores the need for government authorities and policymakers
422 to adopt region-specific strategies in their housing and homelessness policies. In areas
423 with lower rents, policies should place a stronger emphasis on housing subsidies or rent
424 control to alleviate the financial burden on low-income households.

425 Figure 8 (f) shows the association between mortgage affordability and homeles-
426 ness risk. With the increasing proportion of the population under mortgage pressure,
427 its correlation with homelessness risk rises. Specifically, when the proportion of the
428 population under mortgage pressure is below 15%, the correlation between the two
429 variables is at a lower level (around 0.15). However, when it exceeds 15%, the correla-

430 tion between them increases more rapidly, reaching around 0.2. As the proportion of
431 the population under mortgage pressure increases, the correlation with homelessness
432 risk also rises. Specifically, when the proportion of the population under mortgage
433 pressure is below 15 %, the correlation between mortgage pressure and homelessness
434 risk is relatively low (around 0.15). This suggests that with less mortgage pressure,
435 fewer people face economic hardship, reducing the homelessness risk. However, when
436 the proportion of the population under mortgage pressure exceeds 15 %, the correla-
437 tion between the two variables rapidly increases, reaching around 0.2. This indicates
438 that once the proportion of the population under mortgage pressure exceeds a certain
439 threshold, the correlation with homelessness risk increase significantly.

440 Figure 8 shows the relationship between commuting distance and the homelessness
441 risk. It follows an inverted U-shaped pattern. When commuting distance is relatively
442 short (less than 10 km) or long (greater than 25 km), there is a higher correlation with
443 homelessness risk. However, when commuting distance is at an intermediate level,
444 the correlation between the two variables is smaller. When commuting distance is
445 relatively short (less than 10 kilometers), a higher homelessness risk may be related
446 to the following factors. Firstly, city centers are often hubs of economic activity and
447 employment opportunities, attracting a large number of job opportunities, but they
448 may also have high housing prices and rental costs. This leads to some low-income
449 workers or economically vulnerable groups seeking housing near city centers, but due
450 to the high housing price pressure, they may be unable to afford housing and become
451 homeless. On the other hand, when commuting distance is relatively long (greater
452 than 25 kilometers), the increase in the proportion of homelessness risk may be related
453 to the following factors. People look for more affordable housing options in suburban
454 or peripheral areas far from city centers but also face longer commuting distances.
455 Lengthy commutes can increase economic costs and personal burdens, especially for
456 low-income groups who may not be able to afford high transport costs, putting them
457 at risk of homelessness.

458 4.3.4. *Spatial determinants of homelessness risk*

459 Figure 9 shows the spatial determinant of homelessness risk in Australia, which is
460 the explanatory variable with the strongest association with homelessness risk in each

461 location. Regarding three major cities, Sydney, Melbourne, and Brisbane, the most
462 significant variables associated with homelessness risk are found to be pop density,
463 population, and distance to work. The results show a similar spatial pattern in three
464 cities: from the city center to the suburbs, the most influential variables associated
465 with homelessness risk are changing from popden, diswork, and pop, respectively. This
466 indicates that in the central areas of large metropolises, population density has the
467 strongest correlation with homelessness risk. In the outer city areas, the number of
468 homelessness individuals is more closely related to commuting distance. In urban
469 suburbs, population size plays a crucial role concerning homelessness risk. Regarding
470 three smaller cities, Perth, Adelaide, and Hobart, a different spatial pattern is observed.
471 In the central areas of these cities, commuting distance has the most significant impact
472 on homelessness risk. In other areas and suburbs of the cities, the variables with the
473 strongest association with homelessness risk are population density and population
474 size.

475 *4.3.5. Local bi-variate effects in GPI*

476 Figure 10 shows the interaction matrix in six major cities, which describes the in-
477 teraction effects between each pair of variables to homeless risk. The larger the value,
478 the stronger the interaction between the variables, and the greater the impact on the
479 homeless group. Interaction between rental payment and other variables has an im-
480 portant impact on homelessness risk. The spatial heterogeneity of the interaction of
481 variables can be perceived in the heatmap. The interaction between rental payment
482 and population density has significant impacts on homelessness risk, particularly in
483 major cities such as Sydney, Melbourne, and Brisbane. Higher rental payment coupled
484 with elevated population density exacerbates the risk of homelessness. The intensi-
485 fied demand for housing in densely populated areas amplifies rental costs, placing a
486 disproportionate burden on individuals and families with limited financial resources.
487 In Sydney, Melbourne, and Brisbane, where population density is relatively high, the
488 cost of housing tends to surge due to the demand-supply dynamics. As a result,
489 the convergence of high rents and dense population further marginalises economically
490 disadvantaged individuals and households, making them vulnerable to homelessness.
491 Moreover, the competitive housing market in densely populated urban areas can limit

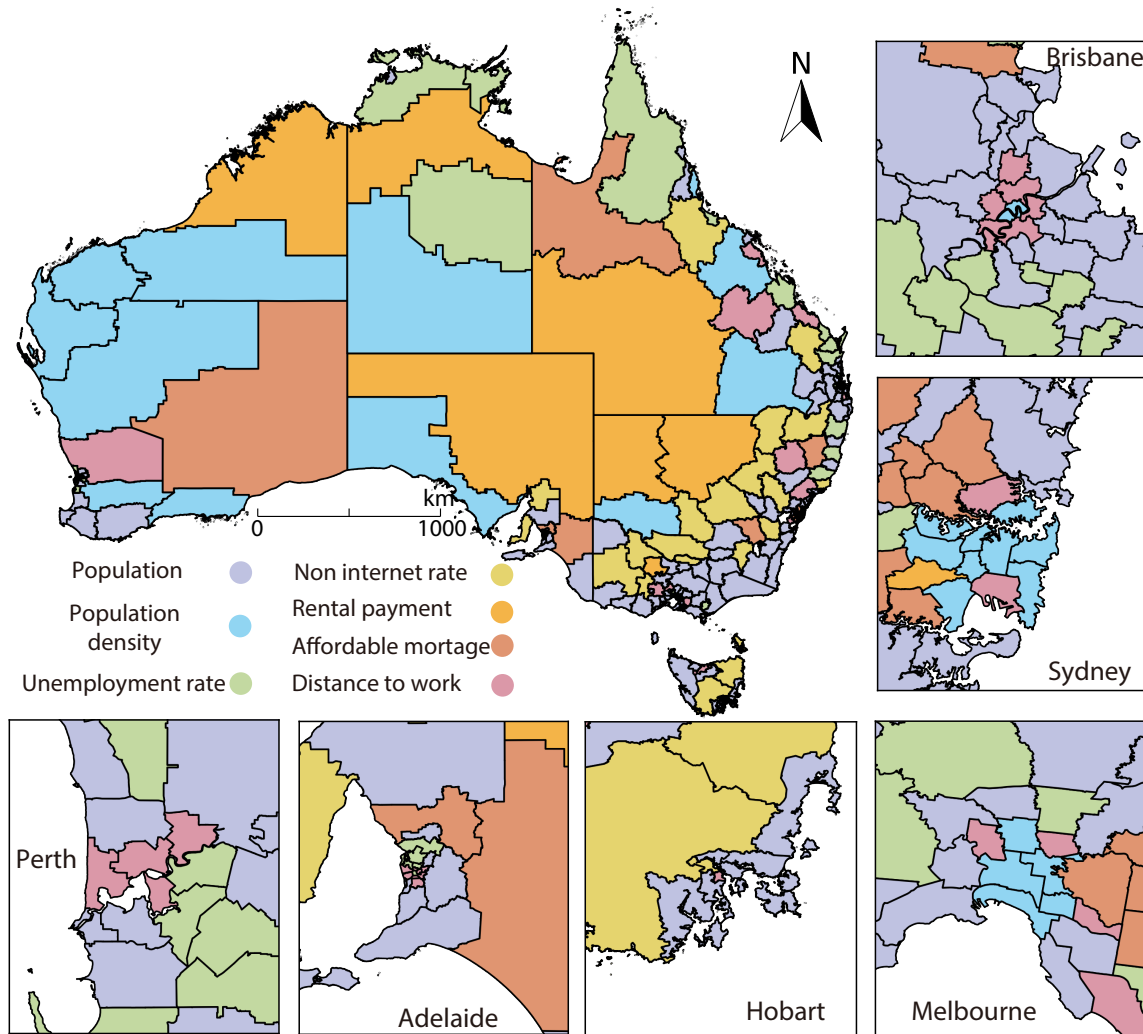


Figure 9: The spatial dominants of the homelessness risk in GPI at each location, which represent the variables that have the strongest association with the homelessness risk.

492 access to affordable housing options, further exacerbating the homelessness risk.

493 The interaction between rental payment and unemployment rate has a significant
 494 impact on homelessness risk in Adelaide and Melbourne. The financial strain caused
 495 by higher rental costs, coupled with limited or no income due to unemployment, cre-
 496 ates a situation where individuals or households become more vulnerable to housing
 497 instability and ultimately homelessness risk. This interaction underscores the critical
 498 relationship between economic factors and homelessness risk.

499 4.4. Model validation

500 In the study, GPI model is validated through a sensitivity analysis (Figure 11).
 501 In 90% of the regions, the percentage change in the dominant variables is below 10%,

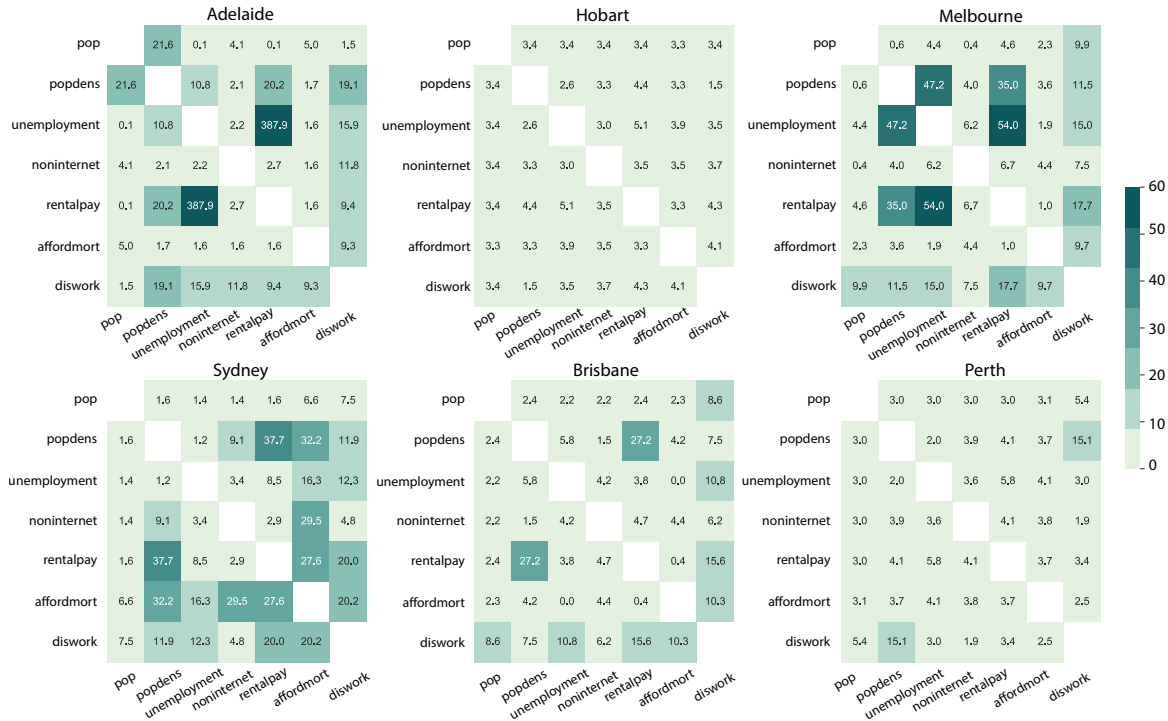


Figure 10: The local interaction effect in GPI.

502 indicating the robustness of our model. The dominant variables exhibiting notable vari-
 503 ations primarily concentrate in northern Australia, an area concurrently characterized
 504 by a relatively high homelessness risk. This may be because that the GPI model relies
 505 on decision tree algorithm for the spatially optimized partitioning of the homelessness
 506 risk, which tends to exhibit sensitivity to outliers to a certain extent.

507 5. Discussion

508 5.1. The advantage of using spatial pattern to explain geographical interactions

509 Previous models for analyzing the association of geographical variables are lim-
 510 ited in exploring the non-linearity of relationships and interactions between multiple
 511 variables. First, they often assume linear relationships between variables, whereas, in
 512 reality, the relationships between geographical variables are often complex and non-
 513 linear (Zhu et al., 2021a). This leads to the existing models needing improvements to
 514 capture the relationships among geographical phenomena accurately. Second, existing
 515 models require to pay more attention to the spatial interactions between geographical
 516 variables. This study proposes the GPI model that uses spatial patterns to character-
 517 ize geographical variables' spatial dependence and spatial heterogeneity for exploring

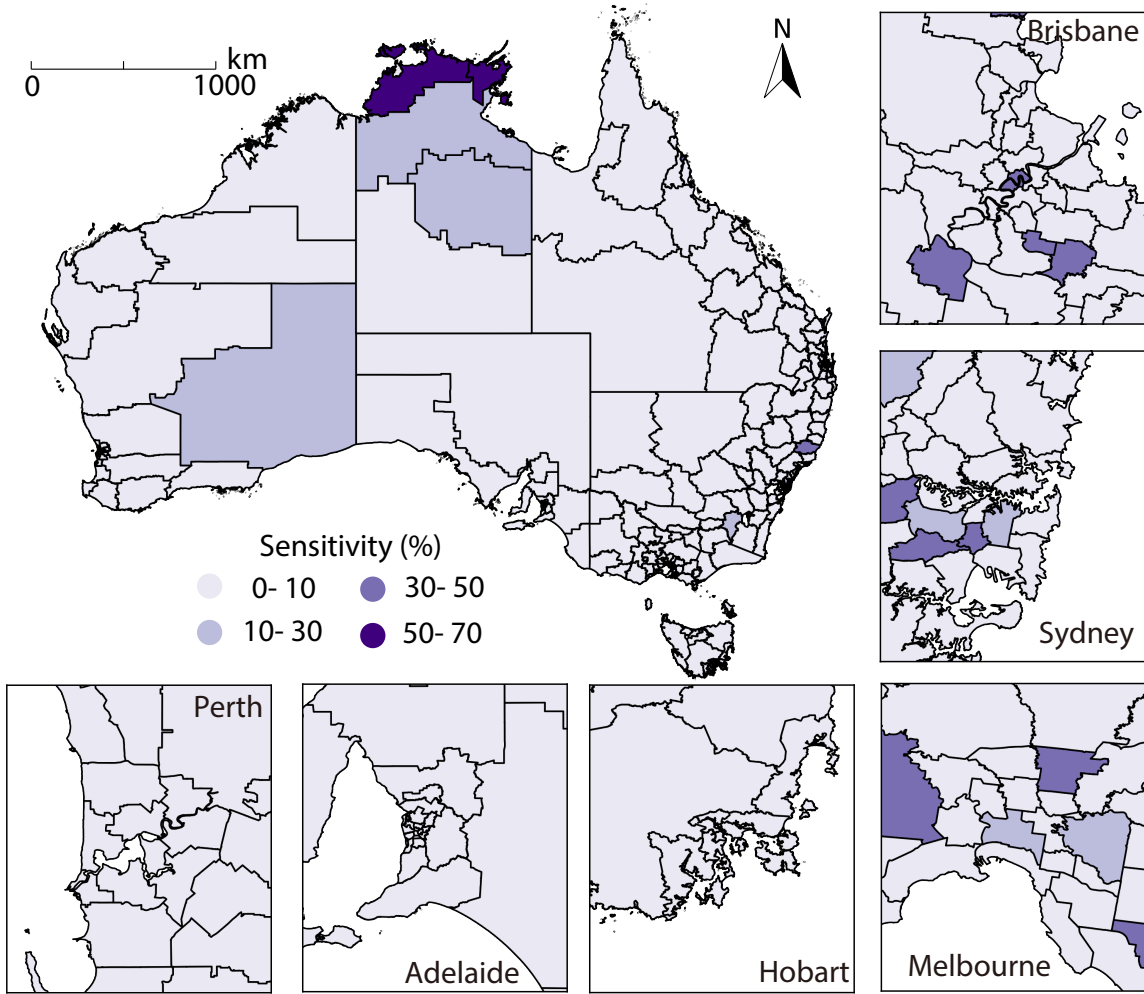


Figure 11: The sensitivity analysis of GPI model. The value refer to the percentage change in the dominant variables when this region is excluded from the model's input data.

518 spatial association. By discretizing the geographical space and analyzing the variance
 519 and mean of response variables for different strata, the GPI model can better describe
 520 the characteristics and patterns of geographical phenomena at different locations, pro-
 521 viding more accurate identification of spatial relationships.

522 The GPI model is based on two fundamental statistical indicators, variance and
 523 mean, to characterize spatial patterns and to describe heterogeneity and stationar-
 524 ity, respectively. Variance is a statistical measure of data dispersion, indicating the
 525 degree of difference between data points and their mean. A higher variance implies
 526 greater differences between data points, representing significant global stratified char-
 527 acteristics. Thus, variance describes global stratified features, i.e., differences between
 528 different groupings. In our model, by calculating the between-group variance of re-

529 sponse variables for different grouping methods, we can assess the contribution of each
530 explanatory variable to the global stratified features. When a specific explanatory
531 variable contributes significantly to the global stratified features, the corresponding
532 grouping method will lead to a higher between-group variance, reflecting the impor-
533 tance and impact of that explanatory variable. Mean is a statistical measure of the
534 central tendency of data, representing the average position of data points around the
535 average value. In our algorithm, we can describe the local stationary features at each
536 location or region by calculating the within-group mean of response variables for each
537 grouping method. When a specific explanatory variable contributes significantly to the
538 local stationary features, the corresponding grouping method will significantly change
539 the within-group mean at that location. Thus, the within-group mean for each location
540 or region can express local stationary features. This study introduces the interpretable
541 machine learning algorithm Shapley to detect the contribution of individual variables
542 in the interaction of multiple explanatory variables. Therefore, the GPI model can ef-
543 fectively identify each variable's contribution to the relationships among geographical
544 phenomena with the consideration of their spatial interactions and provides a more
545 comprehensive explanation of the correlations between geographical variables.

546 In summary, in the GPI model, variance describes global stratified features, re-
547 flecting differences between response variables under different grouping methods. The
548 mean describes local stationary features, reflecting the concentration of response vari-
549 ables around the mean at each location. By computing variance and mean, the GPI
550 model can quantify the contributions of explanatory variables to global and local fea-
551 tures, thereby achieving the interpretation and analysis of spatial patterns.

552 *5.2. Limitations and future works*

553 There are still some limitations to this study, and a few future works are recom-
554 mended. For instance, it is recommended to systematically assess the relationship
555 between the GPI model and existing spatial regression methods. These models have
556 inherent methodological differences, making direct comparisons challenging. In ex-
557 isting spatial regression models, coefficients represent the strength of interaction be-
558 tween independent and dependent variables. For example, in the GWR model, the
559 coefficient represents the impact of a unit change in the independent variable on the

560 dependent variable within a specific geographical area. The proposed GPI model is
561 not constrained by the linearity of relationships but quantifies interaction strength
562 by considering the spatial distribution patterns of geographic variables. We identified
563 an inherent connection between GPI and traditional spatial regression. In GPI, the
564 influence of independent variables on the dependent variable depends on the extent
565 to which their consideration affects the spatial distribution, similar to the concept of
566 coefficients in traditional spatial regression methods.

567 However, there are still paradigmatic differences between GPI and existing spa-
568 tial regression methods, making mutual validation challenging. In existing spatial
569 regression, the relationships between geographic variables are often characterized by a
570 polynomial function, while GPI, based on pattern interaction, resembles more of deci-
571 sion rules (Apté and Weiss, 1997). The advantage of decision rules is their ability to
572 describe discontinuous spatial relationships, where the impact between geographic vari-
573 ables exhibits abrupt, non-continuous changes. In future research, we will discuss the
574 more fundamental connections between GPI and existing spatial regression methods,
575 enabling the design of rational simulation experiments for meaningful inter-method
576 comparisons.

577 Another area for future work would be the potential contribution of GPI with ad-
578 vanced multivariate visualization methods. Given that the basic assumption of GPI
579 is that geospatial variables with similar spatial patterns exhibit stronger relationships,
580 this aligns with human visual understanding of spatial relationships, particularly the
581 ability to recognize spatial correlations based on maps. Hence, it falls within the realm
582 of visual analytics. Currently, multivariate visualization methods are only employed
583 for displaying results. While GPI has demonstrated its capability to reveal complex
584 relationships among geographical variables, integrating advanced multivariate visual-
585 ization methods will further extend the model’s applicability and enhance our insight
586 and understanding of the intricate relationships within geographical data.

587 **6. Conclusion**

588 In this study, we developed a Geographical Pattern Interaction (GPI) model to
589 explore spatial relationships among various geographical variables. The model empha-

590 sizes the spatial patterns of geographical variables under the influence of the interac-
591 tions of explanatory variables for exploring spatial association. By utilizing Spatial
592 Stratified Heterogeneity (SSH) and SHapley Additive exPlanations (SHAP) methods,
593 we quantified spatial associations and interactions within the GPI model. The model
594 effectively identifies spatial associations for individual and multiple variables. Our
595 case study demonstrates the effectiveness of the GPI model in revealing spatial asso-
596 ciations, accommodating spatial interactions, and uncovering non-linear relationships.
597 Overall, the GPI model offers enhanced explanatory power and adaptability, enrich-
598 ing our understanding of complex geographical relationships and providing valuable
599 insights for geographical research and analysis. In future work, cautiously generalizing
600 the GPI model's effectiveness is critical. Moreover, combining GPI with multivariate
601 visualization methods may facilitate a deeper understanding of the spatial patterns
602 and interactions among geographical variables.

603 **Data available statement**

604 The data and codes that support the findings of the present study are available on
605 Figshare at <https://figshare.com/s/1e8136a24509b31ca864>.

606 **Disclosure Statement**

607 No conflict of interest exists in this manuscript, and the manuscript was approved
608 by all authors for publication.

609 **References**

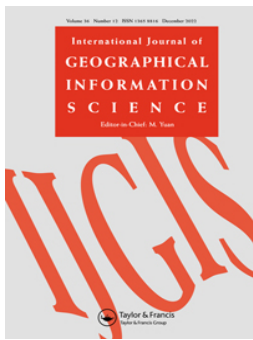
- 610 Anselin, L., 1988. Spatial econometrics: methods and models. volume 4. Springer
611 Science & Business Media.
- 612 Anselin, L., 1989. What is special about spatial data? alternative perspectives on
613 spatial data analysis (89-4). UC Santa Barbara: National Center for Geographic
614 Information and Analysis .
- 615 Anselin, L., 1995. Local indicators of spatial association—lisa. *Geographical analysis*
616 27, 93–115.
- 617 Anselin, L., 2010. Thirty years of spatial econometrics. *Papers in regional science* 89,
618 3–25.
- 619 Anselin, L., Amaral, P., 2023. Endogenous spatial regimes. *Journal of Geographical*
620 *Systems* , 1–26.

- 621 Apté, C., Weiss, S., 1997. Data mining with decision trees and decision rules. *Future*
622 *generation computer systems* 13, 197–210.
- 623 Arbia, G., 2006. *Spatial econometrics: statistical foundations and applications to*
624 *regional convergence*. Springer Science & Business Media.
- 625 Australian Bureau of Statistics, 2016. *Census of population and housing: Estimat-*
626 *ing homelessness*. URL: [https://www.abs.gov.au/statistics/people/housing/](https://www.abs.gov.au/statistics/people/housing/estimating-homelessness-census/2016)
627 [estimating-homelessness-census/2016](https://www.abs.gov.au/statistics/people/housing/estimating-homelessness-census/2016).
- 628 Brunsdon, C., Fotheringham, A.S., Charlton, M.E., 1996. Geographically weighted
629 regression: a method for exploring spatial nonstationarity. *Geographical analysis* 28,
630 281–298.
- 631 Caton, C.L., Dominguez, B., Schanzer, B., Hasin, D.S., Shrout, P.E., Felix, A., Mc-
632 Quiston, H., Opler, L.A., Hsu, E., 2005. Risk factors for long-term homelessness:
633 Findings from a longitudinal study of first-time homeless single adults. *American*
634 *journal of public health* 95, 1753–1759.
- 635 Comber, A.J., Harris, P., Lü, Y., Wu, L., Atkinson, P.M., 2021. The forgotten seman-
636 tics of regression modeling in geography. *Geographical Analysis* 53, 113–134.
- 637 De Marsily, G., Delay, F., Gonçalves, J., Renard, P., Teles, V., Violette, S., 2005.
638 *Dealing with spatial heterogeneity*. *Hydrogeology Journal* 13, 161–183.
- 639 Fotheringham, A.S., Brunsdon, C., Charlton, M., 2000. *Quantitative geography: per-*
640 *spectives on spatial data analysis*. Sage.
- 641 Fotheringham, A.S., Brunsdon, C., Charlton, M., 2003. *Geographically weighted re-*
642 *gression: the analysis of spatially varying relationships*. John Wiley & Sons.
- 643 Fotheringham, A.S., Li, Z., 2023. Measuring the unmeasurable: Models of geographical
644 context. *Annals of the American Association of Geographers* , 1–18.
- 645 Gao, B., Wang, J., Stein, A., Chen, Z., 2022. Causal inference in spatial statistics.
646 *Spatial statistics* 50, 100621.
- 647 Getis, A., Ord, J.K., 1992. The analysis of spatial association by use of distance
648 statistics. *Geographical analysis* 24, 189–206.
- 649 Goodchild, M.F., 2004. The validity and usefulness of laws in geographic information
650 science and geography. *Annals of the Association of American Geographers* 94, 300–
651 303.
- 652 Griffith, D.A., Griffith, D.A., 2003. *Spatial filtering*. Springer.
- 653 Guo, H., Python, A., Liu, Y., 2023. Extending regionalization algorithms to explore
654 spatial process heterogeneity. *International Journal of Geographical Information*
655 *Science* 37, 2319–2344.
- 656 LeSage, J.P., Fischer, M.M., 2008. Spatial growth regressions: model specification,
657 estimation and interpretation. *Spatial Economic Analysis* 3, 275–304.
- 658 Li, Y., Luo, P., Song, Y., Zhang, L., Qu, Y., Hou, Z., 2023. A locally explained
659 heterogeneity model for examining wetland disparity. *International Journal of Digital*
660 *Earth* 16, 4533–4552.
- 661 Li, Z., 2022. Extracting spatial effects from machine learning model using local inter-
662 pretation method: An example of shap and xgboost. *Computers, Environment and*
663 *Urban Systems* 96, 101845.

- 664 Li, Z., 2023. Geoshapley: A game theory approach to measuring spatial effects in
665 machine learning models. [arXiv:2312.03675](https://arxiv.org/abs/2312.03675).
- 666 Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz,
667 R., Himmelfarb, J., Bansal, N., Lee, S.I., 2020. From local explanations to global
668 understanding with explainable ai for trees. *Nature machine intelligence* 2, 56–67.
- 669 Luo, P., Song, Y., 2021. A spatial second-order non-stationary interpolation method
670 for large area mapping. *Abstracts of the ICA* 3, 1–1.
- 671 Luo, P., Song, Y., Huang, X., Ma, H., Liu, J., Yao, Y., Meng, L., 2022a. Identifying de-
672 terminants of spatio-temporal disparities in soil moisture of the northern hemisphere
673 using a geographically optimal zones-based heterogeneity model. *ISPRS Journal of*
674 *Photogrammetry and Remote Sensing* 185, 111–128.
- 675 Luo, P., Song, Y., Wu, P., 2021. Spatial disparities in trade-offs: economic and envi-
676 ronmental impacts of road infrastructure on continental level. *GIScience & Remote*
677 *Sensing* 58, 756–775.
- 678 Luo, P., Song, Y., Zhu, D., Cheng, J., Meng, L., 2022b. A generalized heterogeneity
679 model for spatial interpolation. *International Journal of Geographical Information*
680 *Science* , 1–26.
- 681 Parkinson, S., Batterham, D., Reynolds, M., Wood, G.A., 2019. The changing geogra-
682 phy of homelessness: a spatial analysis from 2001 to 2016. Australian Housing and
683 Urban Research Institute Limited, Melbourne doi:[10.18408/ahuri-5119601](https://doi.org/10.18408/ahuri-5119601).
- 684 Shapley, L.S., et al., 1953. A value for n-person games .
- 685 Song, Y., Wang, J., Ge, Y., Xu, C., 2020. An optimal parameters-based geographical
686 detector model enhances geographic characteristics of explanatory variables for spa-
687 tial heterogeneity analysis: Cases with different types of spatial data. *GIScience &*
688 *Remote Sensing* 57, 593–610.
- 689 Štrumbelj, E., Kononenko, I., 2014. Explaining prediction models and individual pre-
690 dictions with feature contributions. *Knowledge and information systems* 41, 647–665.
- 691 Wang, J., Xu, C., 2017. Geodetector: Principle and prospective. *Acta Geogr. Sin* 72,
692 116–134.
- 693 Zhang, Z., Song, Y., Luo, P., Wu, P., 2023. Geocomplexity explains spatial errors.
694 *International Journal of Geographical Information Science* , 1–21.
- 695 Zhu, D., Liu, Y., Yao, X., Fischer, M.M., 2021a. Spatial regression graph convolutional
696 neural networks: A deep learning paradigm for spatial multivariate distributions.
697 *GeoInformatica* , 1–32.
- 698 Zhu, T., Fonseca De Lima, C.F., De Smet, I., 2021b. The heat is on: how crop
699 growth, development, and yield respond to high temperature. *Journal of Experi-*
700 *mental Botany* 72, 7359–7373.

A6. A generalized heterogeneity model for spatial interpolation

Reference: Luo, P., Song, Y., Zhu, D., Cheng, J., & Meng, L. (2022). A generalized heterogeneity model for spatial interpolation. *International Journal of Geographical Information Science*, 1-26.



A generalized heterogeneity model for spatial interpolation

Peng Luo, Yongze Song, Di Zhu, Junyi Cheng & Liqiu Meng

To cite this article: Peng Luo, Yongze Song, Di Zhu, Junyi Cheng & Liqiu Meng (2022): A generalized heterogeneity model for spatial interpolation, International Journal of Geographical Information Science, DOI: [10.1080/13658816.2022.2147530](https://doi.org/10.1080/13658816.2022.2147530)

To link to this article: <https://doi.org/10.1080/13658816.2022.2147530>



Published online: 20 Nov 2022.



Submit your article to this journal [↗](#)



View related articles [↗](#)



View Crossmark data [↗](#)



RESEARCH ARTICLE



A generalized heterogeneity model for spatial interpolation

Peng Luo^a , Yongze Song^b , Di Zhu^c , Junyi Cheng^d  and Liqiu Meng^a 

^aCartography and Visual Analytics, Technical University of Munich, Munich, Germany; ^bSchool of Design and the Built Environment, Curtin University, Perth, Australia; ^cDepartment of Geography, Environment and Society, University of Minnesota, Minneapolis, MN, USA; ^dInstitute of Remote Sensing and Geographic Information Systems, Peking University, Beijing, China

ABSTRACT

Spatial heterogeneity refers to uneven distributions of geographical variables. Spatial interpolation methods that utilize spatial heterogeneity are sensitive to the way in which spatial heterogeneity is characterized. This study developed a Generalized Heterogeneity Model (GHM) for characterizing local and stratified heterogeneity within variables and to improve interpolation accuracy. GHM first divides a study area into multiple spatial strata according to the sample values and locations of a variable. Then, GHM estimates simultaneously the spatial variations of the variable within and between the spatial strata. Finally, GHM interpolates unbiased estimates and uncertainty at unsampled locations. We demonstrated the GHM by predicting the spatial distributions of marine chlorophyll in Townsville, Queensland, Australia. Results show that GHM improved both the overall interpolation accuracy across the study area and along strata boundaries compared with previous interpolation models. GHM also avoided bull's eye patterns and abrupt changes along strata boundaries. In future studies, GHM has the potential to be integrated with machine learning and advanced algorithms to improve spatial prediction accuracy for studies in broader fields.

ARTICLE HISTORY

Received 15 April 2022
Accepted 9 November 2022

KEYWORDS

Spatial interpolation; spatial heterogeneity; area-to-area kriging; stratified heterogeneity; spatial statistics

1. Introduction

Spatial prediction and interpolation play fundamental roles in geographic analysis (Lam 1983, Mitas and Mitasova 1999, Song 2022). An effective understanding of the characteristics of geographic variables guarantees the accuracy of the spatial interpolation (Oliver and Webster 1990, Zhu *et al.* 2020). Spatial dependence and spatial heterogeneity lay the foundation for spatial interpolations (Goodchild 2004, Tobler 2004). Geostatistical methods employ the spatial dependence of geographical variables for spatial prediction (Matheron 1963, Kyriakidis and Goodchild 2006). In kriging interpolation, widely used in geostatistics and comprising techniques such as ordinary kriging

(OK) and simple kriging, geographic variables are assumed to have second-order spatial stationarity for interpolation (Goovaerts 1997). Kriging assumes that the difference between the values of a geographical variable in two locations is independent of their locations but is only related to their distance (Goovaerts 1997). However, although the spatial second-order stationary assumption is typically satisfied in small areas, it may be weak in large areas with complex surfaces. Previous studies have developed methods to address the spatial non-homogeneity issue in interpolation tasks. For example, kriging with external drift (KED) removes spatial non-homogeneity through continuous drift (Hudson and Wackernagel 1994, Goovaerts 1997, Bourennane *et al.* 2000, Gao *et al.* 2020). However, the spatial stratified non-homogeneity is difficult to eliminate, owing to the continuous drift in the lower order (Chiles and Delfiner 2009).

Spatial non-homogeneity is often manifested by a geographic variable distributed over several spatial strata, each with homogeneous values. Geographical variables often show spatial stratification in the physical world, which is described as spatial stratified heterogeneity (SSH) (Wang *et al.* 2010). The characteristics of SSH make it difficult to construct a stable and reasonable semivariance function across the region (Gao *et al.* 2020). The spatial stratified strategy effectively predicts the spatial distribution at complex surfaces. SSH describes the geographical phenomenon that variable distributes as many homogeneous spatial strata with different spatial means or variances (Song *et al.* 2018, 2020b, Zhang *et al.* 2022). SSH does not require the assumption of spatial second-order stationarity in local heterogeneity. A few recent models have considered SSH to improve spatial interpolation. For example, in the stratified kriging (StK) algorithm, the entire study area is divided into several homogeneous strata, each of which is then subjected to interpolation (Liu *et al.* 2021). However, the numerical information of other strata is completely ignored when interpolating each stratum, and a stratum may only have a limited number of observations after the spatial partition. Ignoring the data between strata leads to limited information for constructing an accurate semivariogram and results in a loss of accuracy. In addition, the spatial division process and subsequent separate interpolation at each stratum may lead to unreasonably sharp changes along the strata boundaries (Gao *et al.* 2020).

Spatial dependence is still present in geographic factors located between the different strata, despite the existence of spatial stratification effect on a large scale (Song and Wu 2021). Geographical differences between strata are usually gradual, and stratification boundaries often manifest themselves as transition areas with certain widths (Fortin *et al.* 1996, De Smith *et al.* 2007, Hutchings *et al.* 2022). Geographic factors in transition areas usually have mixed characteristics with those of neighboring strata. This phenomenon is prevalent in both geographic and socio-economic factors. First, transition areas often exist between different strata of geographic factors, such as elevation and soil moisture. For example, elevation tends to decrease slowly from the plateau to the plain, creating a transition area. Second, spatial dependency between strata is also widely presented in socio-economic factors, such as land use (Preston 1966, Chen *et al.* 2020), nighttime light (Ma *et al.* 2015), economic development level (Erickson 1983) and population density (Luo *et al.* 2019). Significant differences exist between cities at different levels of development and between urban and rural areas (Hutchings *et al.* 2022). However, changes in the boundaries also tend to be gradual.

For example, population density and socio-economic levels tend to decline slowly from urban to rural areas, and there is a mixture of urban and rural characteristics at the urban-rural border. In summary, at larger scales, the distribution of geographical variables is spatially stratified, but there are usually continuous and gradual strata boundaries. However, current spatial interpolation models do not consider this phenomenon in geography.

The motivation of this study is to conduct accurate and reliable spatial prediction for large-scale geographic environments, considering both the existence of spatial stratification and spatial dependence at strata boundaries. A practical solution is to borrow information from other strata to consider both the spatial stratification strategy to ensure overall accuracy and reasonable estimates at the strata boundaries. Specifically, when performing spatial interpolation for strata boundaries, it is essential to consider information from different strata simultaneously. For example, when interpolating the population density in an urban-rural transition area, both urban and rural areas provide the necessary information. When interpolating elevations in the plains-plateau transition area, it is necessary to consider that the elevation in this area has a mixed characteristic of plateaus and plains.

With this motivation, two key issues need to be considered: the identification of transition areas or boundary areas, and the method used to borrow information from different areas. However, only a few studies have considered information borrowing to interpolate, and no study has considered the region in which borrowed information is needed. For example, a point mean of the surface with stratified non-homogeneity (P-MSN) algorithm was proposed to conduct interpolation in a large marine area (Gao *et al.* 2020). The study area was divided into several strata, and the semivariogram between each pair of strata was estimated using OK. It does not consider the existence of transition areas between partitions and assumes that the contribution of information from other regions to interpolation is offset by interference.

In summary, large-area stratified interpolation requires the process of bringing information from other strata, but this process often introduces high uncertainty in the result and leads to substantial computational cost. Therefore, trade-offs exist in the amount of information obtained from the outside stratum. The main concern is that observations from other strata or remote areas can introduce noise. We assume $n + k$ observations in the study area, including n observations in the interpolated stratum and k observations from other m strata. Previous studies have controlled the trade-offs by arranging different weights for the n observations within the stratum and k observations outside the stratum. This study provides a new method to automatically borrow information from other strata without manually adjusting the weights of different strata. The basic idea is to merge observations from each other strata separately and fit the semivariogram between two parts: the observations in the interpolated stratum and outside strata. Although all the observations from the outside strata are used to solve the spatial dependence between different strata, each stratum provides only one value in the fitted semivariogram. Thus, the uncertainty from the outside strata is expected to be limited when the information is borrowed. In addition, this approach reduces computational consumption and improves the interpolation efficiency.

Calculating the weights of other strata when conducting spatial interpolation was an important task in this study. Areal interpolation algorithms, such as area-to-point kriging (ATAP) and area-to-area kriging (ATAK) (Sadahiro 2000, Kyriakidis 2004, Goovaerts 2010), were developed to estimate the weights of areas. These algorithms were proposed to handle the interpolation of data at different scales and were used to disaggregate areal data into spatial prediction at the levels of points and different areas (Guan *et al.* 2011, Geddes *et al.* 2013, Hu and Huang 2020). Given its effectiveness in representing the spatial association between different areas, it is reasonable to believe that ATAK can be used to calculate the weights of other strata and characterize the spatial association between different strata.

In this study, a Generalized Heterogeneity Model (GHM) was developed. It combines ATAK and OK for the interpolation of spatial second-order non-homogeneity areas with high accuracy and efficiency. A specific geographical variable that presents spatial stratified non-homogeneity in a complex surface is distributed over many spatially homogeneous strata. Geographical variables that describe the same region exhibit spatial dependence, whereas variables that describe different regions exhibit spatial heterogeneity. The relationship between observations from different strata is represented by the relationship between strata. Thus, ATAK was introduced to characterize the spatial dependence between different strata and construct the corresponding semivariogram. In this way, information is borrowed while maintaining spatial dependence inside the homogeneous stratum. In addition, most of the information from other strata is noisy and interferes with interpolation accuracy. Using ATAK to characterize the spatial dependence between different strata may address this problem, because only the average value of each outside strata is considered in building the semivariogram.

We demonstrated the GHM using spatial interpolation of marine chlorophyll in Townsville, Queensland, Australia. Reliable and spatially continuous data on marine environments are essential for the conservation of biodiversity. However, in most marine areas, only sparse and unevenly distributed point samples are available, which is particularly pronounced in Australian marine regions (Li and Heap 2008). Therefore, it is critical to develop effective interpolation models for marine environments (Elumalai *et al.* 2017). Spatial interpolation in marine environments is challenging for two reasons. First, the spatial second-order stationary assumption is easily violated in large-area marine environments because of the highly dynamic movement of water masses and the resulting uneven distribution of ocean components (Gao *et al.* 2015, 2020). Stratification has been found and verified in marine environments (Bowman and Esaias 1981). Effective spatial interpolation technology that considers SSH is necessary for marine research. Second, spatial interpolation is a relatively difficult task in the marine environment, compared to that for the land environment, because of the lack of supporting explanatory variables. Without any supporting data, the mapping performance relies on understanding the characteristics of geographic variables through reasonable interpolation algorithms, which is an ideal case to verify the advantages of the proposed GHM.

The accuracy and effectiveness of GHM were evaluated through cross-validation and comparisons with previous related interpolation models, including OK, KED and

StK. The remainder of this paper is organized as follows. Section 2 describes the whole process of GHM for interpolation. Section 3 presents the implementation of GHM for the interpolation of marine chlorophyll in Townsville, Queensland, Australia. Section 4 discusses the findings and research contributions, and the study is concluded in Section 5.

2. Generalized heterogeneity model (GHM)

In this study, a Generalized Heterogeneity Model (GHM) was proposed to conduct stratified spatial prediction while considering information from other strata. This section is formulated as follows: concepts of GHM, development of the objective function, process of solving the function, optimal neighboring search strategy and execution of the GHM.

2.1. Concepts of GHM

Figure 1 shows the differences among classical geostatistical interpolation algorithms. The geographical data were assumed to be distributed as lower on the left and higher on the right. The interpolation theory of OK and KED is primarily based on spatial dependence, constructing semivariance functions at a global level. StK considers the existence of SSH by partitioning the space and constructing separate semivariance functions in each stratum to improve the accuracy. P-MSN considers that information

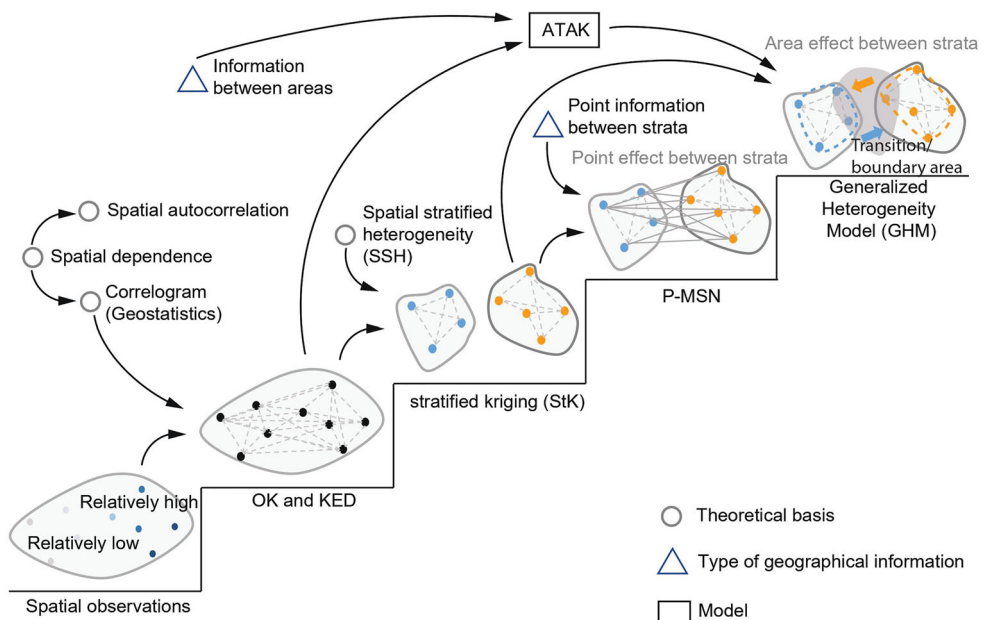


Figure 1. Theoretical basis of the Generalized Heterogeneity Model (GHM) and the relevant models: OK (ordinary kriging), KED (kriging with external drift), StK (stratified kriging), P-MSN (point mean of surface with stratified non-homogeneity) and ATAK (area-to-area kriging).

between different regions is borrowed from each other using OK to construct a semi-variance function of the point level between strata.

GHM has two theoretical innovations compared to previous studies: (1) GHM considers the existence of strata boundaries (i.e. transition areas between strata) and the spatial dependence of strata at the boundaries. The borrowed information is used to improve the interpolation in these regions; (2) GHM borrows information from other strata in the form of an area. This has the promise of introducing valid information while avoiding interference information from other strata as much as possible.

2.2. Objective functions of GHM

Given that a spatially stratified area is divided into several homogeneous strata, the interpolated value is the weighted sum of two parts: observations within the interpolated stratum and observations outside the interpolated area. Assuming that spatial division has already been conducted, and there exist several homogeneous strata, the interpolated value is calculated as follows:

$$\hat{Z}_0 = Z_{in} + Z_{out} = \sum_{i=1}^n \lambda_i Z_i + \sum_{j=n+1}^{n+k} \lambda_j Z_j \quad (1)$$

where \hat{Z}_0 is the interpolated value, Z_{in} is the weighted sum of the observations in the interpolated stratum, and Z_{out} is the weighted sum of the observations in the other strata. n is the number of observations in the interpolated stratum, and k is the number of observations in the other strata. Z_i and Z_j are the observation values, where λ_i and λ_j are the weights of the observations.

Weight vector λ includes the weights of all observations, which are characterized as follows:

$$\lambda = [\lambda_{in}, \lambda_{out}] = [\lambda_1, \lambda_2, \lambda_3 \dots \lambda_n, \lambda_{n+1}, \dots \lambda_{n+k}] \quad (2)$$

where λ_{in} is the weight vector of the observations in the interpolated stratum, consisting of $\lambda_1, \lambda_2, \lambda_3 \dots \lambda_n$. λ_{out} is the weight vector of the observations in the other strata, consisting of $\lambda_{n+1}, \dots \lambda_{n+k}$.

The interpolated values are estimated using the solved weight vector. Similar to other geostatistic models, two objective functions should be developed to obtain the best linear unbiased estimation:

$$\begin{cases} E(\hat{Z}_0 - Z_0) = 0 \\ \min \text{Var}(\hat{Z}_0 - Z_0) \end{cases} \quad (3)$$

By introducing the Lagrange multiplier, the two formulas are transformed into the following determinants (Appendices A and B):

$$\begin{bmatrix} R_{1,1} & \dots & R_{1,n+k} & m_{s1} \\ R_{2,1} & \dots & R_{2,n+k} & m_{s1} \\ \dots & & & \dots \\ R_{n+k,1} & \dots & R_{n+k,n+k} & m_{s2} \\ m_{s1} & \dots & m_{s2} & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \dots \\ \lambda_{n+k} \\ L \end{bmatrix} = \begin{bmatrix} R_{1,0} \\ R_{2,0} \\ \dots \\ R_{n+k,0} \\ m_{s1} \end{bmatrix} \quad (4)$$

where $R_{i,j}$ is the covariance between available observations i and j ($i, j = 1, \dots, n+k$).

$R_{i,0}$ is the covariance between interpolated point 0 and available observation i ($i = 1, \dots, n+k$), where λ_i is the weight of the i th observation. m_{s_1} , m_{s_2} are the expectations of the variables inside and outside the interpolation stratum, respectively.

2.3. Solution

The matrix in the determinant (Equation (4)) describes three types of spatial dependence: dependence of observations in the interpolated stratum, dependence of observations between the interpolated stratum and the other strata and dependence of observations in the other strata. In geostatistical analysis, spatial dependence is described using a semivariogram.

We defined two kinds of semivariograms: the within-semivariogram S_w and the between-semivariogram S_b . S_w represents the spatial dependence of the observations in the interpolated stratum. S_b describes the spatial dependence between the observations in different strata, regardless of whether the observations in the interpolated stratum are included. In the equation for the determinant (Equation (4)), S_w includes the covariance of all the pairs of observations in the interpolated stratum. S_w includes the covariance between the two parts: observations in the interpolated stratum, and observations in the other strata. The S_w of each stratum was calculated using OK (Figure 2(c,d)).

S_b was solved by introducing ATAK. ATAK was initially used for interpolation using polygon data. To build a semivariogram between polygons, the polygons are disaggregated into points. The semivariogram between each pair of points is calculated and regarded as a semivariogram between polygons (Gotway and Young 2002, Yoo and Kyriakidis 2006). For example, in ATAK, the predictor of an area with an unknown value is calculated using a linear combination of covariances between nearby areas. The calculation of S_b in a stratum is the most important part of the GHM. First, all observations are merged into different areas (Figure 2(e,f)). Each observation in the

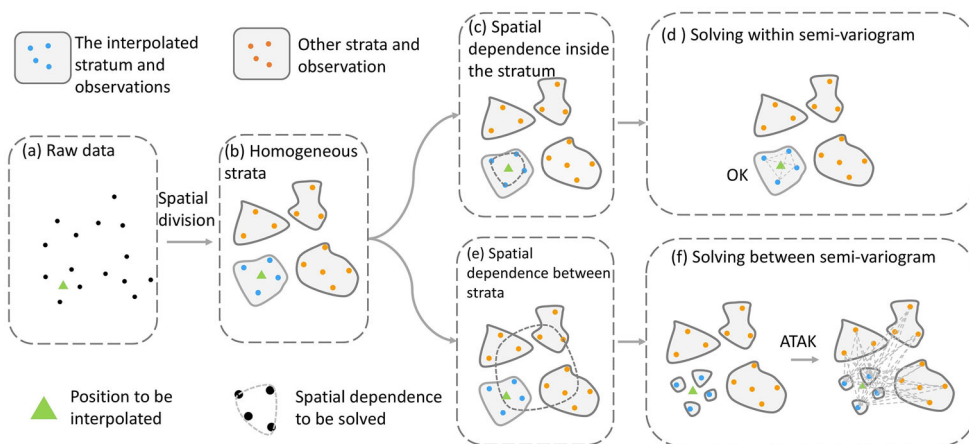


Figure 2. Semantic figure of the GHM for interpolation, including the spatial division, solving of the within-semivariogram S_w using OK, and the solving of the between-semivariogram S_b using ATAK.

interpolated stratum is regarded as an area with only one observation. Observations in other strata are separately merged into their respective strata. Second, the semivariance between the two areas is calculated using ATA_K as follows:

$$R(v_i, v_j) = R(a_m, a_k) = \frac{1}{\sum_{s=1}^{P_m} \sum_{t=1}^{P_k} (w_s * w_t)} \sum_{s=1}^{P_m} \sum_{t=1}^{P_k} (w_s * w_t) R(u_s, u_t) \quad (5)$$

where v_i and v_j are the two observations, a_m is the stratum where v_i occurs, and a_k is the stratum where v_j occurs. If v_i is the observation from the interpolated stratum, then it is equal to a_m . u_s and u_t are the observations of a_m and a_k , respectively; P_m and P_k are the numbers of observations of a_m and a_k , respectively and w_s and w_t are the weights of u_s and u_t , respectively, which are usually equal to one. u_s and u_t are necessary to estimate the area from discretized points.

It should be mentioned that although the observations from other strata are merged into several areas (i.e. strata), all observations are used to solve the spatial dependence between the different strata. Thus, the fit between semivariograms considers the spatial dependence in the interpolated stratum as much as possible and borrows information from the other strata.

2.4. Optimal neighboring search strategy

In geostatistical models, only the number of nearest observations (N_{max}) or observations within a certain range are used for interpolation, considering the computing efficiency (Lichtenstern 2013). As shown in Figure 3(a), only locations near the strata boundaries have neighboring observations from other strata and borrow information from other strata. In this study, we define an observation that may have neighboring observations from another stratum as a boundary observation, and if this condition does not hold, it is a non-boundary observation.

The N_{max} in the non-boundary area only controls information from the same stratum because only observations in the interpolated stratum are used for interpolation (Figure 3(a)):

$$\hat{Z}_{0_{nb}} = Z_{in} = \sum_{i=1}^{N_{max}} \lambda_i Z_i \quad (6)$$

where $\hat{Z}_{0_{nb}}$ is the interpolated value in the non-boundary area.

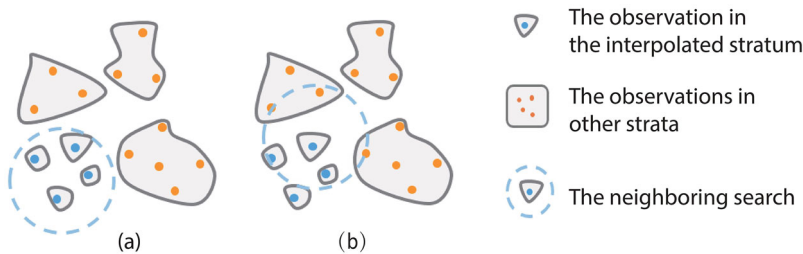


Figure 3. Influence of the neighboring search range to GHM: (a) observations that do not include borrowed information from other strata and (b) observations taking into account information from other strata.

In contrast, N_{max} in the boundary area determines how much information is borrowed from other strata (Figure 3(b)), because some neighboring observations are from other strata:

$$\hat{Z}_{0_b} = Z_{in} + Z_{out} = \sum_{i=1}^n \lambda_i Z_i + \sum_{j=n+1}^{N_{max}} \lambda_j Z_j \quad (7)$$

where \hat{Z}_{0_b} is the interpolated value in the boundary area.

Therefore, different search ranges (e.g., the number of nearest observations for interpolation) should be considered in the boundary and non-boundary areas. It is necessary to separately optimize N_{max} in the two areas. Optimal neighboring search strategy for boundary and non-boundary observations is proposed in this study. First, the boundary area of the interpolated stratum is identified (Figure 4(a)). For each stratum, the observations inside are divided into boundary area observations (Figure 4(a), dark color) and non-boundary area observations (Figure 4(a), light color). There are many methods for identifying boundary observations, eg edge detection for remote sensing images and buffer analysis of boundary lines. In addition, for sample point data, boundary identification is conducted depending on the number of neighboring observations from other strata or the distance to other strata. Second, after identifying the strata boundaries, the optimization of N_{max} in the boundary area (Figure 4(b)) and non-boundary area (Figure 4(c)) is executed. Different N_{max} values are set in the boundary and non-boundary regions; then, GHM interpolation is executed to obtain

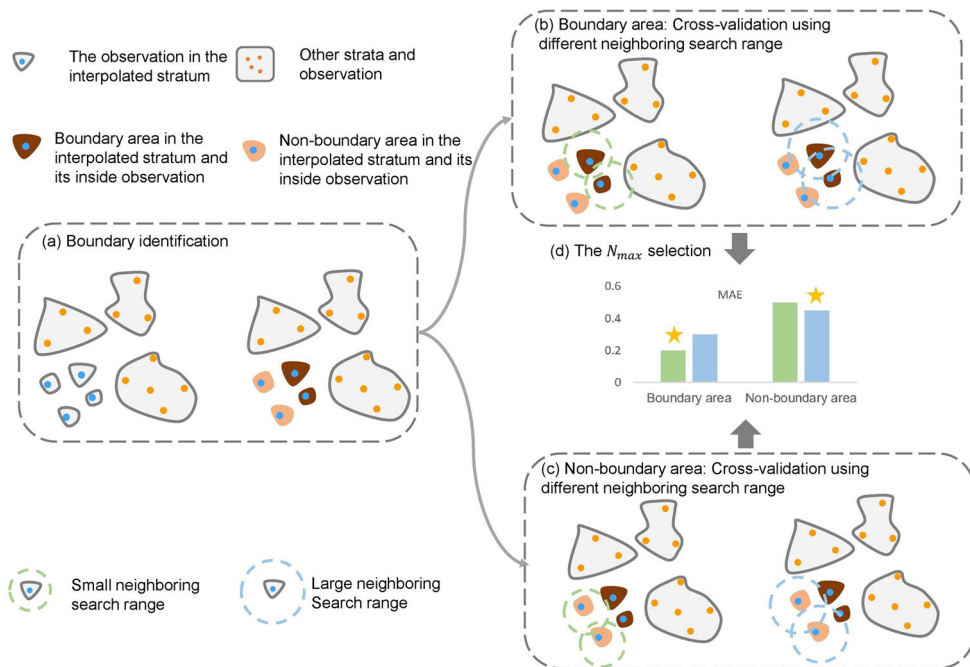


Figure 4. N_{max} selection process: (a) identify the boundary area of the interpolated stratum; cross-validation using different neighboring search ranges in (b) boundary areas and (c) non-boundary areas and (d) selection of N_{max} . The N_{max} with the highest validation accuracy is selected and labeled with a star.

the interpolation accuracy using cross-validation. Finally, the N_{max} values in the boundary and non-boundary areas with the highest accuracy are selected as the final N_{max} values for GHM interpolation (Figure 4(d)).

2.5. Execution process of GHM

The execution process of interpolation using GHM is summarized as follows. First, a large area, which is spatial second-order non-stationary, is divided into several homogeneous strata. The division process is conducted based on administrative units, geographical grids or expert experience and using clustering and image segmentation algorithms (Likas *et al.* 2003, Gao *et al.* 2020). Second, the semivariograms for each stratum are fitted. The variogram inside the stratum was fitted using OK. The variogram between different strata was fitted using ATAK.

Third, N_{max} optimization was conducted for each stratum. Finally, interpolation was conducted for each stratum. The interpolated value for the locations in the boundary area is the weighted sum of the neighboring observations inside and outside the stratum. The interpolated values in locations at non-boundary areas are the weighted sum values of the neighboring observations inside the interpolated stratum.

3. Case study: mapping marine chlorophyll using GHM

3.1. Study area and data

In this case study, we demonstrated GHM by spatial interpolation of marine chlorophyll in Townsville, Queensland, Australia. Marine chlorophyll data in the study area, including 4136 observations, were collected by the Australian National Facility for Ocean Gliders on 1 August 2010, which is a part of the Integrated Marine Observing System (IMOS) (Davies *et al.* 2018). The IMOS ocean observing mission is focused on the Australian coast and is critical for understanding the north-south transport of freshwater, heat and biogeochemical properties. These data are collected by sensors containing environmental information, such as temperature, chlorophyll, salinity and turbidity at different locations and instrument depths. The chlorophyll content ranges from 0.01 to 311.13, with an average value of 0.62, and the standard deviation is 5.12. Figure 5 shows the location of the study area and the spatial distribution of the marine chlorophyll observations in the study area. Figure 5 shows that a significant number of observations are located in very close proximity, considering that there are 4137 observations but only a few hundred that can be visually distinguished. Therefore, it is necessary to perform declustering prior to spatial modeling.

3.2. GHM-based interpolation

In this case, samples of marine chlorophyll observations are the only data used for spatial prediction. It is difficult to collect explanatory variables to support this prediction. Thus, the GHM provides an opportunity to accurately predict the spatial distribution of marine chlorophyll in the study area. The GHM-based interpolation of marine chlorophyll includes the following six steps (Figure 6): data pre-processing, spatial

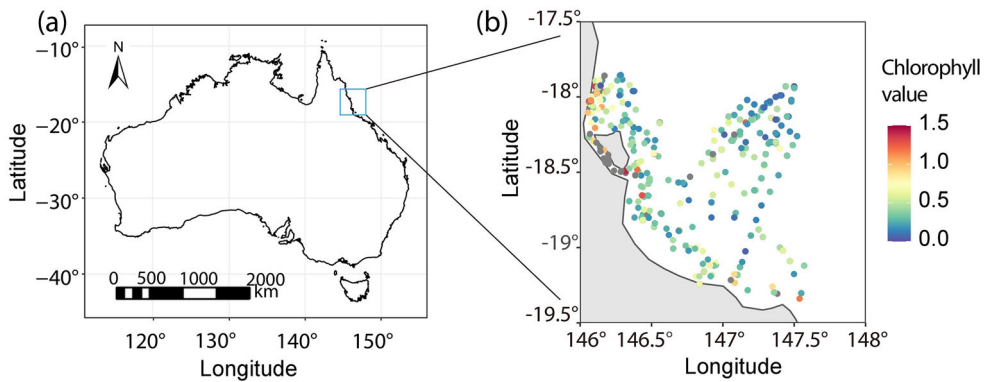


Figure 5. (a) Study area and (b) spatial distribution of marine chlorophyll samples in the study area. Observations with chlorophyll values out of the color legend (0 to 1.5) are shown as grey dots.

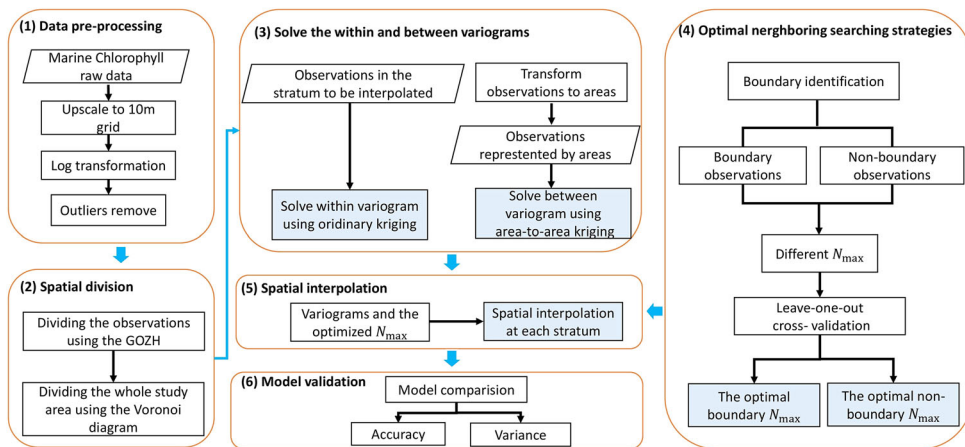


Figure 6. Flowchart of interpolation using GHM.

division, semivariogram solving, N_{max} optimization, spatial interpolation and accuracy assessment. These steps are introduced in the following paragraphs.

The first step is data processing. All observations were up-scaled to 10m grids using an average composite. Then a log transformation was conducted because the original data approximately followed a log-normal distribution. Outliers were removed by eliminating any observations that were more than twice the standard deviation from the mean.

Second, a spatial division was conducted. An ideal partitioning result should have the smallest intra-partition variance and the largest inter-partition variance. Thus, a geographically optimal zones-based heterogeneity (GOZH) model was used to divide the entire area into several homogeneous strata (Wang *et al.* 2010, 2016, Song *et al.* 2020a, Luo *et al.* 2021, 2022). The GOZH model is a SSH model that allows spatial division that considers the maximum homogeneity within each stratum. Spatial

division is regarded as an optimization task in the GOZH model and is formulated as follows:

$$\Omega = \text{Max} \left[1 - \frac{SSW}{SST} \right] \quad (8)$$

where Ω is a measure of the spatial stratified heterogeneity, SSW is the sum-of-squares within the stratum, and SST is the sum-of-squares total of marine chlorophyll in the whole study area. In the GOZH model, Ω was solved step-wise, with the same optimization objective, and was used to split the process of the Classification and Regression Tree method (Chipman *et al.* 1998, Luo *et al.* 2022). Spatial division was conducted after Ω was determined.

The spatial division guided by GOZH maintains spatial homogeneity inside each stratum as much as possible. Thus, the large spatial second-order non-homogeneous area was divided into several homogeneous strata. During the spatial division process, longitude and latitude were the two explanatory variables for marine chlorophyll. The entire study area was divided into several strata according to longitude and latitude using the GOZH model.

After the spatial division of the observations, a spatial division in the area without observations was conducted. In this study, we created a Voronoi diagram, using all the observations, in which the area of each diagram belongs to the same stratum as the corresponding observation. It should be noted that the GOZH-based spatial discretization method is not compulsive to be used in GHM. The optimal spatial discretization method should be selected according to the research question and corresponding expert knowledge. We chose the GOZH and Voronoi diagrams because they are intuitive and straightforward, obtaining the greatest non-homogeneity between different strata.

Third, the within- and between-semivariogram in each stratum was solved using OK and ATAK, respectively. For each stratum, the within-semivariogram S_w describes the spatial autocorrelation of all observations. The R package 'gstat' was used to build the semivariogram. The S_w varied with the strata. For a specific stratum, all observations inside were transformed into an area with a uniform value, and all other strata were merged into an area. Then, ATAK was used to construct a semivariogram between these areas. The R package 'atakrig' was used to conduct ATAK.

Fourth, the boundary observations were identified according to the number of nearest observations at other strata, and the optimal N_{max} was selected based on cross-validation. For a particular stratum, we counted the N nearest observations around each observation. The proportion of N observations from the other strata was then counted. We set a series of N values and chose 15 as the optimal value based on visual inspection, ensuring that the derived boundary area had a reasonable number of observations and a stable line structure. Hence, 15 neighboring observations were calculated for each observation, and the boundary observation was identified if at least one neighboring observation was from other strata. After identifying the boundary areas, the optimal N_{max} values for the boundary areas and non-boundary areas were identified. To select the optimal parameter, the range of N_{max} for the GHM is from 10 to 15. Because the largest N_{max} is smaller than 15, the identified non-boundary observation would have no neighboring observations from other strata. For each

stratum's boundary or non-boundary area, we set a different N_{max} and then performed GHM interpolation to verify and calculate the interpolation accuracy. Finally, we selected the N_{max} with the highest accuracy as the final interpolation parameter. In this study, leave-one-out cross-validation was used to optimize the parameters. Leave-one-out cross-validation is a particular case of k-fold cross-validation, in which the number of folds equals the number of observations (Wong 2015). Leave-one-out cross-validation is widely used to assess the interpolation performance of geostatistical models (Gong *et al.* 2014). Each observation was selected as the test set individually, and interpolation at this location was conducted using all other observations. In this study, the mean absolute error (MAE) derived from the leave-one-out cross-validation was used to compare the interpolation accuracy in the boundary and non-boundary areas to select the optimal N_{max} .

Fifth, interpolation was conducted within each stratum using the solved variograms and optimized N_{max} . Finally, the performance of GHM was evaluated using the leave-one-out cross-validation by comparison with three related geostatistical models, OK, KED and StK, which were conducted using the R package 'gstat'. In this study, StK shared the same spatial division result as GHM to fairly compare the performance. The semivariograms in OK and KED were solved using the R package 'gstat'. StK shared the same semivariograms at each stratum with the within-semivariogram of GHM. The N_{max} values for OK, KED and StK were selected according to sensitivity analysis. It should be mentioned that there was only one N_{max} in the entire study area for the three models, for both the boundary and non-boundary areas.

3.3. Results

3.3.1. Data pre-processing and neighboring search optimization

This section presents the results of data pre-processing, spatial division and N_{max} optimization. Figure 7(a,b) shows the process of spatial division using the GOZH model. The study area was divided into three strata, considering the highest homogeneity of marine chlorophyll within each stratum. The boundary identification results are shown in Figure 7(d). Most boundary observations were located in the regions from -18° to -18.5° .

Figure 8 shows the process of N_{max} optimization for the three strata. In stratum A, the MAE in the non-boundary area was higher than that in the boundary area. The highest boundary MAE, at 0.635, corresponded to an N_{max} of 10. The boundary MAE decreased with an increase in N_{max} and reached its lowest value when N_{max} was 14. However, the MAE in the non-boundary area was very stable, ranging from 0.460 to 0.462. The lowest value of 0.460 was obtained when the N_{max} was 14. Compared with stratum A, there was no significant pattern of N_{max} in stratum B. The boundary MAE had the lowest value (0.435) when the N_{max} was 13. The non-boundary MAE had the lowest value (0.372) when the N_{max} was 11. In stratum C, the MAE in non-boundary was far higher than that in the boundary area, ranging from 0.516 to 0.529. Here, the lowest non-boundary MAE (0.516) corresponded to an N_{max} of 14. The boundary MAE increased with an increase in N_{max} from 10 to 15 and decreased slightly when N_{max} was 16. The lowest value, which is 0.268, corresponded to an N_{max} of 10.

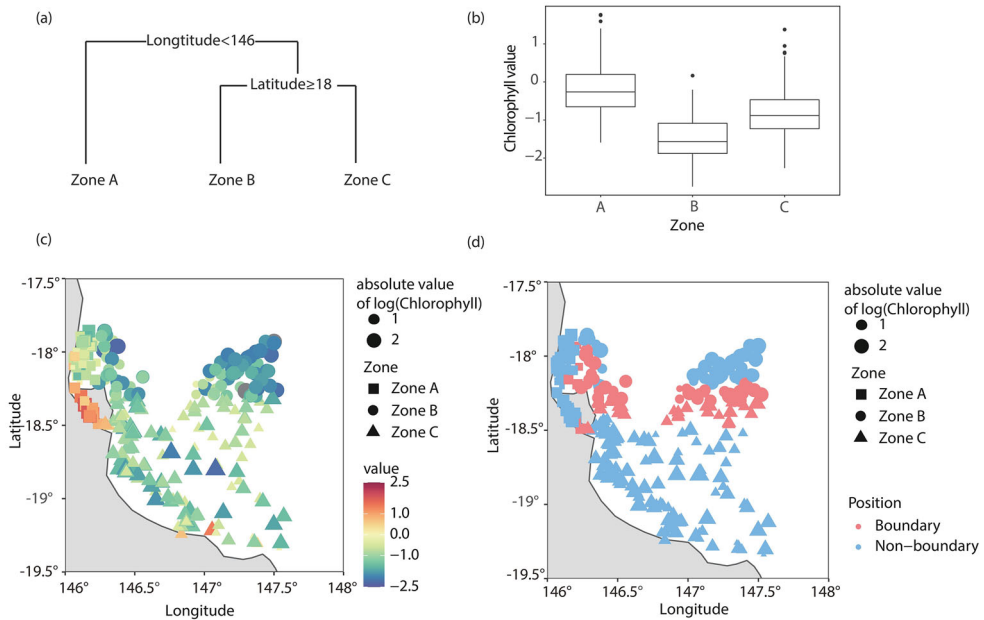


Figure 7. Data pre-processing and spatial division: (a) spatial division process based on the GOZH model; (b) box plots of marine chlorophyll data distribution in the divided strata; (c) observations belonging to three divided strata and (d) observations belonging to boundary and non-boundary areas.

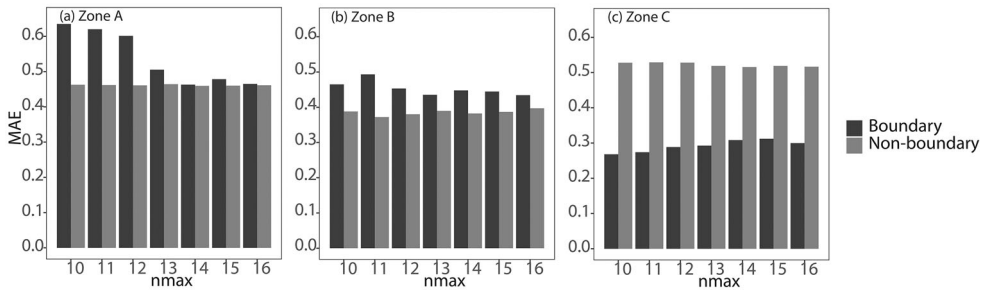


Figure 8. Selection of the optimal neighboring samples. N_{max} is the number of the nearest observations used for interpolation: (a) stratum A; (b) stratum B and (c) stratum C.

Table 1 lists the geostatistical parameters of the four models. The variogram of the original value was used by OK, and the variogram without drift was used by StK. For a specific stratum, the within-semivariogram S_w and between-semivariogram S_b characterize the spatial dependence within the interpolated stratum and between different strata, respectively. These two variograms were used by the GHM. In addition, the within-semivariogram S_w was also used by StK in each stratum.

3.3.2. Accuracy assessment and interpolation results

Table 2 shows the accuracy of the four models in the entire area, boundary area and non-boundary area. The two stratified models, StK and GHM, had better interpolation performance than the non-stratified models OK and KED. GHM had the highest

Table 1. Variogram of marine chlorophyll data under different conditions: variogram of the original value (OK), variogram without drift (StK), between-variogram (GHM) at each stratum, and within-variogram (GHM and StK) at each stratum.

Area	Type of variogram	Model	Sill	Nugget	Range (km)
Whole Area	Variogram of the original value	Sph	0.38	0.20	3.64
	Variogram without drift	Exp	0.35	0.00	3.30
Stratum A	Within-variogram	Sph	0.27	0.00	0.66
	Between-variogram	Sph	0.29	0.00	1.39
Stratum B	Within-variogram	Sph	0.23	0.00	1.41
	Between-variogram	Sph	0.30	0.00	12.95
Stratum C	Within-variogram	Sph	0.32	0.00	0.95
	Between-variogram	Sph	0.40	0.00	11.30

Table 2. Comparison of interpolation models, including OK, KED, StK and GHM, based on cross-validation.

Model	All area RMSE			Boundary area			Non-boundary		
	MAE	RMSE	R ²	MAE	RMSE	R ²	MAE	RMSE	R ²
OK	0.485	0.626	0.392	0.415	0.563	0.440	0.501	0.640	0.376
KED	0.481	0.622	0.398	0.407	0.557	0.450	0.498	0.636	0.382
StK	0.465	0.600	0.424	0.380	0.509	0.531	0.485	0.620	0.397
GHM	0.457	0.589	0.439	0.374	0.505	0.536	0.476	0.607	0.414

accuracy among the four models, with the lowest MAE and root-mean-square (RMSE) values. The MAE values for the whole area, boundary area and non-boundary area were 0.457, 0.374 and 0.476, respectively. The MAE of the GHM for the entire area was 6.1%, 5.3% and 1.7% lower than OK, KED and StK, respectively. The RMSE of the GHM for the entire area was 6.3%, 5.6% and 1.9% lower than OK, KED and StK, respectively. In addition, GHM performed better interpolation in both the boundary area and non-boundary area. The MAE in the boundary and non-boundary areas for StK was 1.6% and 1.9% higher than that of GHM, respectively. GHM takes into account information from the other strata in the boundary, so the accuracy significantly increased, showing that marine chlorophyll in boundaries between strata has a spatial dependency, leading to smooth change. Borrowing information between different strata is necessary to improve the interpolation accuracy. The accuracy in non-boundary areas is increased owing to the use of the optimal parameter for the search area in the GHM. The accuracy of KED was slightly higher than that of OK in terms of lower RMSE and MAE. The MAE of KED was 0.83% lower than that for OK for the entire area. It performed better in the boundary area than in the non-boundary area, with an MAE 2.0% lower than that for OK.

Figure 9 shows the MAE in the boundary (Figure 9(a)) and non-boundary areas (Figure 9(b)) in the three strata. The GHM produces interpolation with the highest accuracy in all strata, especially in the boundary areas. The results showed that the stratified interpolation model significantly improved the accuracy of the boundary area. StK and GHM had lower MAE values than OK and KED. In the boundary area of stratum A, the interpolation MAE of GHM was 0.463, which was 8.4%, 12.5% and 13.0% lower than that of StK, KED and OK, respectively. In the boundary areas of strata B and C, the accuracies of GHM and StK were similar but were significantly higher

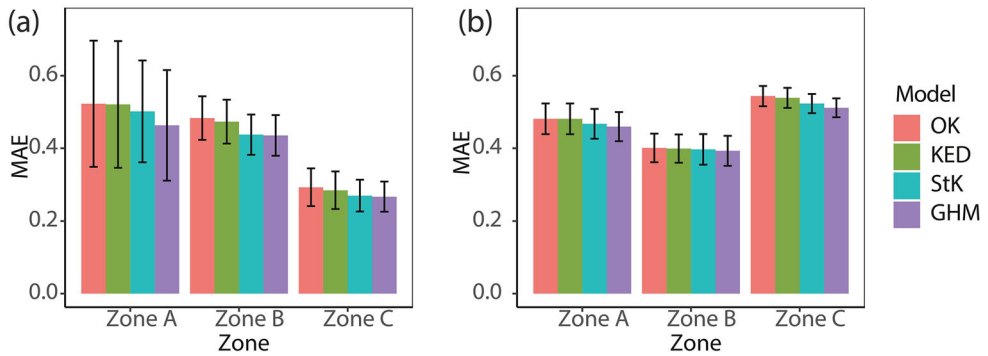


Figure 9. Comparison of the cross-validation MAE in different strata in the study area: (a) boundary area and (b) non-boundary area.

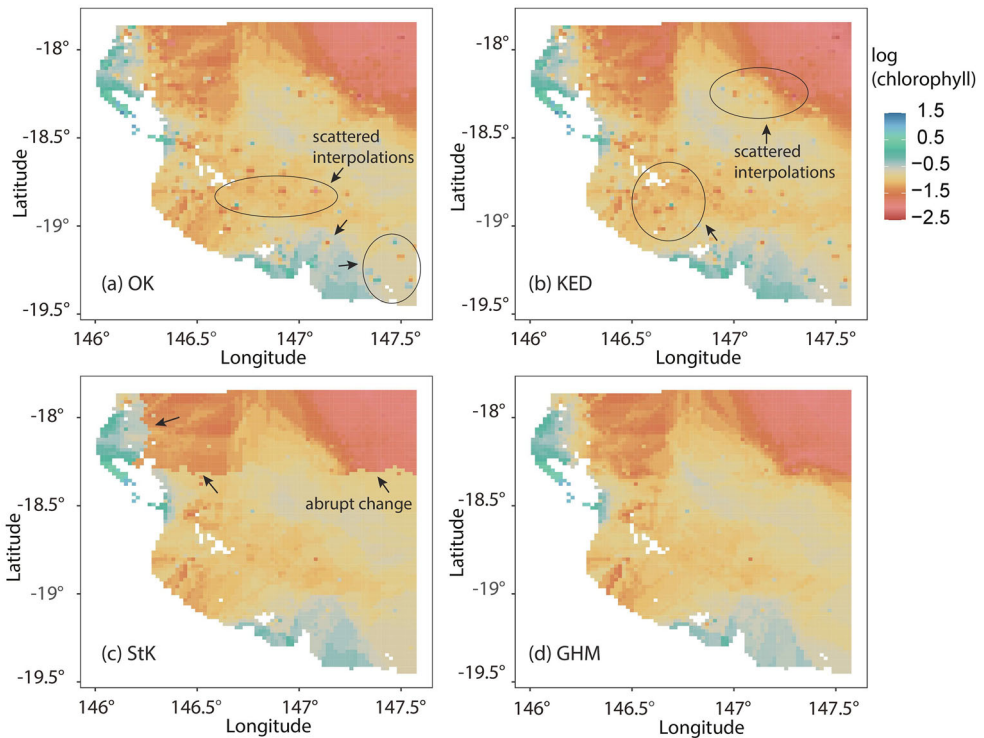


Figure 10. Spatial interpolation results of four models: (a) OK; (b) KED; (c) StK and (d) GHM.

than those of OK and KED. In the non-boundary area, MAE was still slightly lower than that of the other three models.

Figure 10 shows the interpolation results obtained by OK, KED, StK and GHM. Two non-stratified models, OK and KED, had smooth interpolation results because the study area was regarded as a whole, and only one semivariogram was built for the two models. However, the spatial prediction contained bull's eye patterns around the samples. The interpolation result from StK avoided the bull's eye patterns but showed abrupt changes along the boundaries between the strata. StK conducted the

interpolation in each region separately, and no information was borrowed from the other regions. Another stratified model, GHM, had a smooth result along the boundary, which was similar to the results of OK and KED. Ocean chlorophyll usually has a smooth distribution; therefore, the continuous change along the boundary is reasonable. In addition, the GHM avoided bull's eye patterns. In summary, the results demonstrate that our proposed GHM had the highest accuracy in both boundary area and non-boundary area and avoided bull's eye patterns and abrupt changes along the boundaries, enabling more reasonable spatial interpolations.

3.3.3. Interpolation uncertainty analysis

Figure 11 shows the spatial distribution of the estimation error from the GHM (Figure 11(a)) and the difference in absolute error between the GHM and the other three models (Figure 11(b–d)). As shown in Figure 11(b–d), GHM performed the best

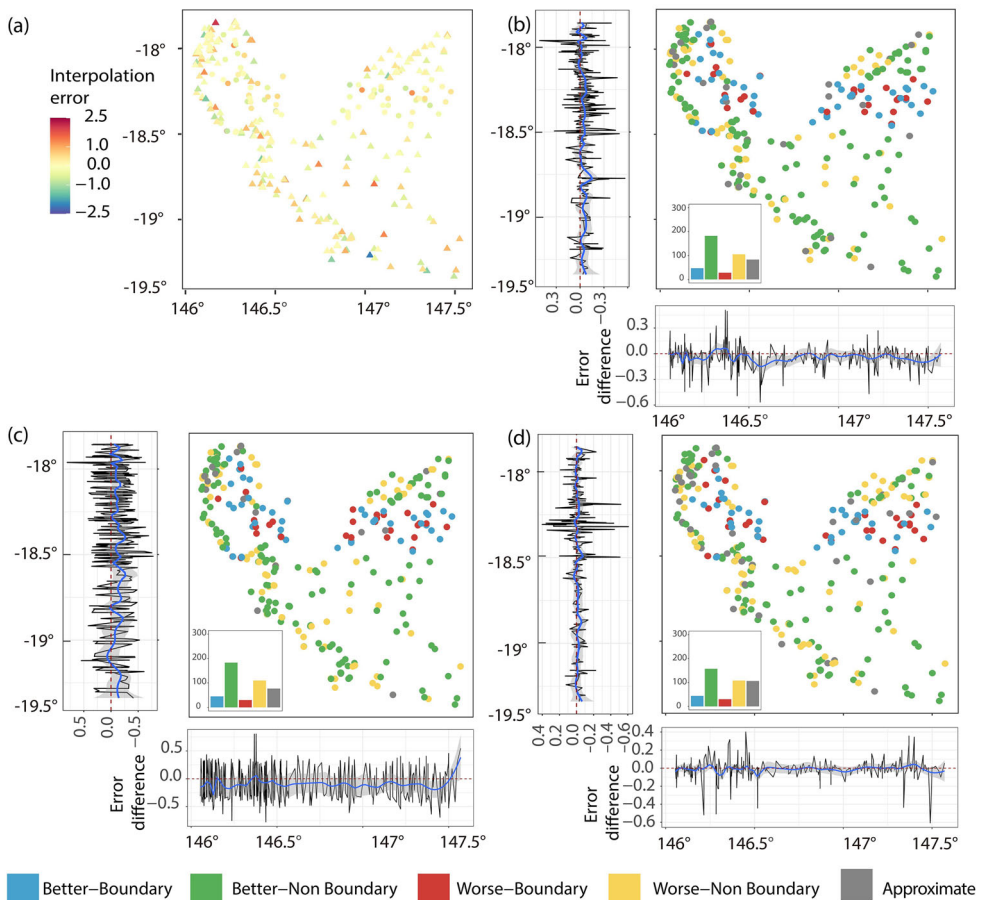


Figure 11. (a) Spatial distributions of the estimated errors of GHM, and the error difference between GHM and other models: (b) OK, (c) KED, (d) StK. An error difference lower than -0.05 means that GHM has better performance, and an error difference greater than 0.05 means the compared model has better performance. An absolute error difference within 0.05 shows that GHM has a performance similar to the compared model.

estimation (blue and green) among the four interpolation algorithms in most observations. Figure 11(b) shows a comparison of GHM and OK. GHM achieved better results at 51.5% of the observations than the other models. For the accuracy in the cross section, the estimation accuracy of GHM was higher than that of OK, except near 146.3°E . The average accuracy was higher in all the regions. The division between regions A and B occurs near 146.3°E . This indicates that the stratified process loses accuracy at the boundary between the two regions. From the longitudinal section, the average accuracy of GHM was higher than that of OK in most of the longitudinal cross sections. Figure 11(c) shows a comparison between GHM and KED. The average accuracy of the GHM was higher than that of the KED in most of the longitude and latitude cross sections.

Figure 11(d) shows the comparison of uncertainty for GHM and StK. Although the accuracy of GHM was higher than that of StK in the vast majority of the observed points, the absolute difference between the two estimation accuracies was not significant. The accuracy difference curves were around the value of zero in both the longitude and latitude cross sections. However, the uncertainty difference values were lower than zero in the majority of the regions, indicating that GHM had a relatively higher accuracy. It is worth mentioning that the accuracy of GHM was higher than that of StK at the boundary, between the latitudes -18°N and -18.5°N . This proves that the information-borrowing strategy of GHM was essential for reducing interpolation uncertainty.

The estimation variance is an essential indicator of interpolation performance. Figure 12(a) shows the variance in the observations derived from the GHM using

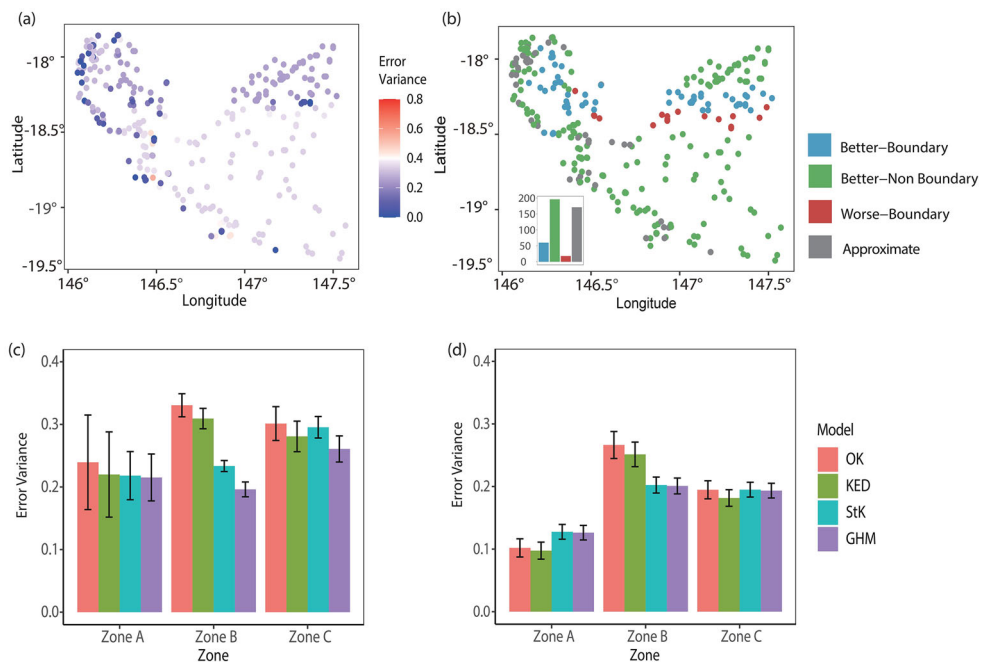


Figure 12. Cross-validation estimation variance of interpolations: (a) variance of GHM; (b) variance comparison between GHM and StK and variance in (c) boundary and (d) non-boundary areas in four models.

leave-one-out cross-validation. The variance ranged from 0 to 0.8. We compared the variance difference between the two stratified interpolation models, the proposed GHM and StK (Figure 12(b)). The results show that the GHM had a lower variance in most observations. GHM had a lower variance at 78% of boundary observations and 100% of non-boundary observations. A comparison of the error variance among the four models at different strata is shown in Figure 12(c,d). Generally, the two non-stratified models showed a higher average error variance than StK and GHM. Exceptions were the non-boundary areas of strata A and C. GHM showed the lowest average variance in most areas, including all boundary areas. The average variance of GHM (0.19) was significantly lower than that of the other models in the boundary area of stratum B, which was 42%, 39% and 17% lower than OK (0.33), KED (0.31) and StK (0.23), respectively. Two non-stratified models, OK and KED, had a lower variance in the non-boundary area of stratum A.

The estimation variance from the final interpolation process was mapped to the entire study area (Figure 13). Considering the estimation variance derived from leave-one-out cross-validation, the two stratified models have lower error variance than the non-stratified models. In the OK and KED models, the error variance at locations near observations was significantly lower than that at locations in other areas. In some areas with the highest estimation variance, such as the southwestern study area, GHM and StK also had a relatively low variance compared with OK and KED. GHM and StK had similar error variances in the non-boundary region because their interpolation

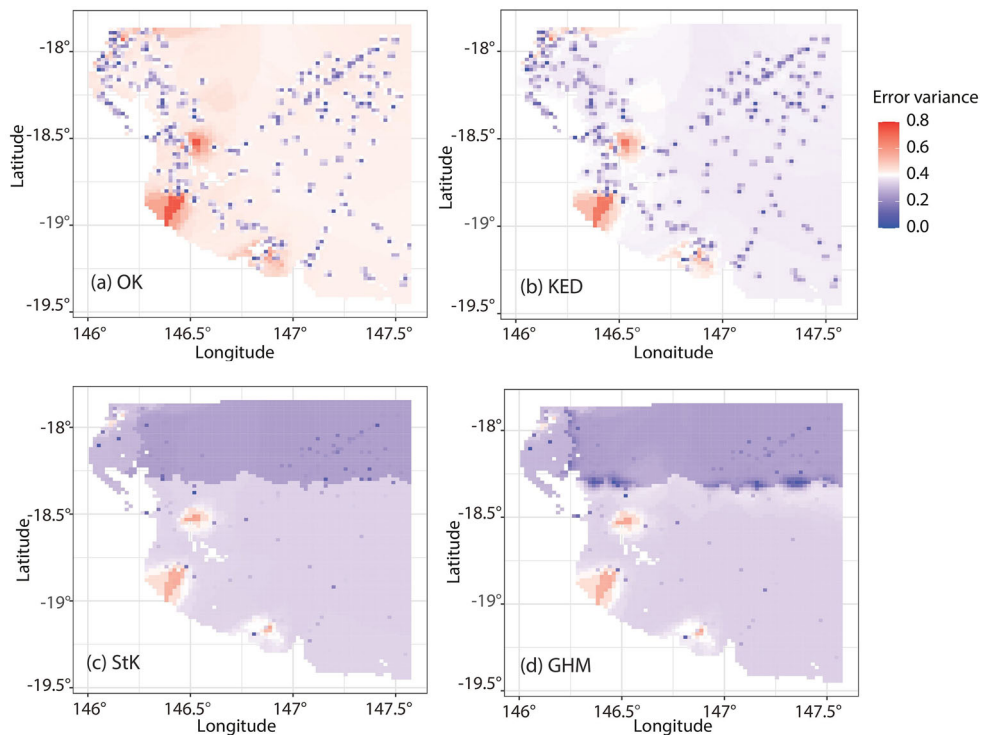


Figure 13. Spatial distributions of interpolation error variance derived from (a) OK, (b) KED, (c) StK and (d) GHM.

processes are quite similar apart from the boundary areas. GHM estimates interpolations with lower variance along the boundary than StK, because the boundaries are characterized and information from other strata is used in the GHM. However, in a remote area of the boundary region in stratum C, GHM had a higher interpolation variance than StK. The neighboring searching strategy arranges different N_{max} values to the boundary areas. The optimized N_{max} is effective for improving the overall estimation accuracy but may increase the uncertainty in the region close to the non-boundary area.

In summary, the results show that GHM has the highest estimation accuracy in terms of MAE and RMSE. GHM-based interpolations also had a lower variance, especially along the boundary regions. This demonstrates the effectiveness of the GHM and the necessity of borrowing information for the stratified geostatistical model.

4. Discussion

Spatial prediction is a challenging task for geostatistical models given that spatial second-order stationarity may be violated. Geographical variables tend to involve spatial stratification, with homogeneity within the stratification. Although heterogeneity exists between different strata, the stratification boundaries within geographical variables are bounded by spatial dependencies.

Previous studies have explored stratified interpolation algorithms, such as dividing the study area into homogeneous strata and removing continuous drift. However, the stratification process leads to information loss which limits the interpolation accuracy. Several methods have been developed to conduct the stratified interpolation while borrowing information from different strata. However, these methods ignore the spatial dependence that exists in the transition area between regions. In addition, when solving the kriging objective function, the constraints of these methods are typically too strong. In this study, we propose a GHM for interpolation in a spatially non-homogeneous large area. OK and ATAK were used to characterize the spatial dependence inside the interpolated stratum and between strata, respectively. The study area was divided into strata that were second-order stationary prior to interpolation. To interpolate each stratum, the semivariogram within the observations was solved using OK, and the semivariogram between observations from different strata was solved using ATAK. In addition, the boundaries between different strata were identified. The optimal neighboring observations (N_{max}) in the boundary and non-boundary areas was estimated using leave-one-out cross-validation.

In this study, we demonstrated the GHM through spatial prediction of marine chlorophyll in a study area in Australia. In similar cases, it is difficult to collect explanatory variables to support spatial prediction, and GHM performs well for spatial prediction. The results showed that the GHM had the highest interpolation accuracy in terms of RMSE and MAE. We found that the stratification strategy effectively improved interpolation accuracy in a large area with spatial second-order non-stationarity. Two stratified models, StK and our proposed GHM, had higher accuracies than OK and KED. They also had similar interpolation results in non-boundary areas. The GHM performed with a higher accuracy in the boundary area than StK. In addition, the interpolation

result from StK exhibited a sharp change along the boundary, resulting from spatial division. The GHM had a smoother estimation along the boundary because it borrowed information from other strata. A comparison of the error variances from the four models also verifies the necessity of information borrowing. The GHM had a lower estimation variance along the boundary than the StK, reducing the interpolation uncertainty. Apart from the three baseline models, we also compared the interpolation performance between the GHM and another information-borrowing model, the P-MSN. The results show that the MAE and RMSE of the P-MSN in the study area were 0.465 and 0.600, respectively. Its performance was similar to that of StK but lower than that of GHM. In addition, the MAE of GHM was 3.5% and 1.5% lower than that of P-MSN in the boundary and non-boundary areas, respectively. This indicates that the interpolation performance of the GHM is generally better than that of P-MSN, and the improvement is most evident in the boundary areas.

The main contributions of this study are summarized as follows: First, an effective and practical method for large-area mapping was developed by combining OK and ATAK. ATAK was used to characterize the spatial dependence between homogeneous strata. The introduction of ATAK is highly effective in large-area mapping. Second, an optimal neighboring search strategy was introduced to better borrow information from other strata when constructing the between-semivariogram S_b . Third, the results indicated that the influence of other homogeneous strata might be characterized as an area effect.

Spatial second-order stationarity is challenging in large areas, and a single semivariogram cannot reflect the real spatial dependence of the entire area. Dividing the region into several homogeneous small regions is a straightforward way to improve interpolation accuracy, which is the basic idea of StK. However, StK may lead to each stratum having limited observations, which introduces the difficulty of fitting a semivariogram. In addition, conducting interpolation at different strata leads to a sharp change along the stratum boundary. In some special environments, such as cliffs and faults, environmental variables may exhibit sharp changes. In such a situation, spatial division is conducted according to special geography, and the sharp change in interpolation results from StK might be reasonable. However, most geographical variables change gradually in the real world, especially marine environmental variables such as the chlorophyll selected in our case study. This sharp change is unreasonable for the spatial distribution of most geographical variables. Borrowing information from other strata along the boundary is the basic idea of the GHM, and its effectiveness was verified by our results.

However, there are still some limitations to this study. First, the methods used for the spatial division and optimization of N_{max} should be improved. Spatial division was conducted using the GOZH model, which divides the study area according to the longitude and latitude. The division process may introduce uncertainty because the geographical environment usually does not have a spatial pattern similar to that of the longitude and latitude. In this study, our primary aim was to propose and verify the idea of borrowing information using ATAK; therefore, only simple and straightforward methods were used for these steps. More advanced and accurate spatial division algorithms should be used in future studies, such as k-means and density-based spatial

clustering of applications with noise (DBSCAN) (Hartigan and Wong 1979, Hahsler *et al.* 2019). Second, the proposed method is a geostatistic interpolation model without combining machine learning and other learning methods. As previous work has proven the potential of machine learning in spatial interpolation (Zhu *et al.* 2020), we will explore how to combine it with GHM to obtain better accuracy.

5. Conclusions

In this study, a Generalized Heterogeneity Model (GHM) was developed to improve the spatial interpolation accuracy of data in large areas. The study space was divided into several strata according to geographical distributions. The spatial dependence within observations in the interpolated stratum was characterized by OK, and the spatial dependence between observations from different strata was characterized by ATAK. The results of the case study demonstrate that GHM had the highest accuracy in terms of MAE and RMSE, compared with other widely used interpolation models, including OK, KED and StK. In addition, the GHM avoided bull's eye patterns and abrupt changes along the strata boundaries.

This paper presents an effective approach for interpolating spatial second-order non-stationary surfaces. We characterized the spatial dependence of different heterogeneous partitions by introducing ATAK, which we hope will inspire future spatial interpolation and prediction. For large-scale regions, both natural and socio-economic variables tend to exhibit spatial second-order non-stationarity, and the GHM method has the potential to effectively interpolate and predict them spatially.

Acknowledgements

We would like to thank the editors and anonymous reviewers for their constructive suggestions and comments for improving this manuscript. We also thank Dr. Maogui Hu for his insightful suggestions.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Peng Luo is currently a PhD candidate at the Chair of Cartography and Visual Analytics at the Technical University of Munich, Germany. His research interests include spatial association modeling, social sensing and applied artificial intelligence.

Yongze Song is a Lecturer at Curtin University, Australia and a Fellow of the Royal Geographical Society (with IBG), United Kingdom. His current research interests include geospatial analysis methods, spatial statistics, sustainable development and infrastructure management.

Di Zhu is an Assistant Professor of Geographic Information Science at the University of Minnesota, Twin Cities (UMN). His research aims at generating both theoretical and actionable insights from spatiotemporal data by exploring the frontiers that bridge geospatial analysis, artificial intelligence and social sensing.

Junyi Cheng is currently a PhD candidate at the Institute of Remote Sensing and Geographic Information System, Peking University. His research interests include data mining, the application of big data in public security and information systems.

Liqiu Meng is a professor of Cartography at the Technical University of Munich and a member of German National Academy of Sciences. She is serving as Vice President of the International Cartographic Association. Her research interests include geodata integration, mobile map services, multimodal navigation algorithms, geovisual analytics and ethical concerns in social sensing.

ORCID

Peng Luo  <http://orcid.org/0000-0002-3680-8509>
 Yongze Song  <http://orcid.org/0000-0003-3420-9622>
 Di Zhu  <http://orcid.org/0000-0002-3237-6032>
 Junyi Cheng  <http://orcid.org/0000-0002-9225-2824>
 Liqiu Meng  <http://orcid.org/0000-0001-8787-3418>

Data and codes availability statement

The data and codes that support the findings of the present study are available on Figshare at <https://doi.org/10.6084/m9.figshare.19604245>.

References

- Bourennane, H., King, D., and Couturier, A., 2000. Comparison of kriging with external drift and simple linear regression for predicting soil horizon thickness with different sample densities. *Geoderma*, 97 (3-4), 255–271.
- Bowman, M.J., and Esaias, W.E., 1981. Fronts, stratification, and mixing in long island and block island sounds. *Journal of Geophysical Research*, 86 (C5), 4260–4264.
- Chen, K., et al., 2020. Land use transitions and urban-rural integrated development: theoretical framework and china's evidence. *Land Use Policy*, 92, 104465.
- Chiles, J.P., and Delfiner, P., 2009. *Geostatistics: modeling spatial uncertainty*. Vol. 497. New York, NY: John Wiley & Sons.
- Chipman, H.A., George, E.I., and McCulloch, R.E., 1998. Bayesian cart model search. *Journal of the American Statistical Association*, 93 (443), 935–948.
- Davies, C.H., et al., 2018. A database of chlorophyll a in Australian waters. *Scientific Data*, 5 (1), 1–8.
- De Smith, M.J., Goodchild, M.F., and Longley, P., 2007. *Geospatial analysis: a comprehensive guide to principles, techniques and software tools*. Leicester, UK: Troubador Publishing Ltd.
- Elumalai, V., et al., 2017. Spatial interpolation methods and geostatistics for mapping ground-water contamination in a coastal area. *Environmental Science and Pollution Research International*, 24 (12), 11601–11617.
- Erickson, R.A., 1983. The evolution of the suburban space economy. *Urban Geography*, 4, 95–121.
- Fortin, M.-J., et al., 1996. Quantification of the spatial co-occurrences of ecological boundaries. *Oikos*, 77 (1), 51–60.
- Gao, B., et al., 2015. A stratified optimization method for a multivariate marine environmental monitoring network in the Yangtze River estuary and its adjacent sea. *International Journal of Geographical Information Science*, 29 (8), 1332–1349.
- Gao, B., et al., 2020. Spatial interpolation of marine environment data using P-MSN. *International Journal of Geographical Information Science*, 34 (3), 577–603.

- Geddes, A., *et al.*, 2013. Stochastic model-based methods for handling uncertainty in areal interpolation. *International Journal of Geographical Information Science*, 27 (4), 785–803.
- Gong, G., Mattevada, S., and O'Bryant, S.E., 2014. Comparison of the accuracy of kriging and IDW interpolations in estimating groundwater arsenic concentrations in Texas. *Environmental Research*, 130, 59–69.
- Goodchild, M.F., 2004. The validity and usefulness of laws in geographic information science and geography. *Annals of the Association of American Geographers*, 94 (2), 300–303.
- Goovaerts, P., 1997. *Geostatistics for natural resources evaluation*. New York, NY: Oxford University Press on Demand.
- Goovaerts, P., 2010. Combining areal and point data in geostatistical interpolation: applications to soil science and medical geography. *Mathematical Geosciences*, 42 (5), 535–554.
- Gotway, C.A., and Young, L.J., 2002. Combining incompatible spatial data. *Journal of the American Statistical Association*, 97 (458), 632–648.
- Guan, Q., Kyriakidis, P.C., and Goodchild, M.F., 2011. A parallel computing approach to fast geostatistical areal interpolation. *International Journal of Geographical Information Science*, 25 (8), 1241–1267.
- Hahsler, M., Piekenbrock, M., and Doran, D., 2019. dbscan: fast density-based clustering with r. *Journal of Statistical Software*, 91 (1), 1–30.
- Hartigan, J.A., and Wong, M.A., 1979. Algorithm as 136: a k-means clustering algorithm. *Applied Statistics*, 28 (1), 100–108.
- Hu, M., and Huang, Y., 2020. atakrig: an r package for multivariate area-to-area and area-to-point kriging predictions. *Computers & Geosciences*, 139, 104471.
- Hudson, G., and Wackernagel, H., 1994. Mapping temperature using kriging with external drift: theory and an example from Scotland. *International Journal of Climatology*, 14 (1), 77–91.
- Hutchings, P., *et al.*, 2022. Understanding rural–urban transitions in the global south through peri-urban turbulence. *Nature Sustainability*, 5 (11), 1–7.
- Kyriakidis, P.C., 2004. A geostatistical framework for area-to-point spatial interpolation. *Geographical Analysis*, 36 (3), 259–289.
- Kyriakidis, P.C., and Goodchild, M.F., 2006. On the prediction error variance of three common spatial interpolation schemes. *International Journal of Geographical Information Science*, 20 (8), 823–855.
- Lam, N.S.N., 1983. Spatial interpolation methods: a review. *The American Cartographer*, 10 (2), 129–150.
- Li, J., and Heap, A.D., 2008. *A review of spatial interpolation methods for environmental scientists*. Canberra, Australia: Geoscience Australia Canberra.
- Lichtenstern, A., 2013. Kriging methods in spatial statistics. Munich, Germany.
- Likas, A., Vlassis, N., and Verbeek, J.J., 2003. The global k-means clustering algorithm. *Pattern Recognition*, 36 (2), 451–461.
- Liu, Y., *et al.*, 2021. Geographical detector-based stratified regression kriging strategy for mapping soil organic carbon with high spatial heterogeneity. *CATENA*, 196, 104953.
- Luo, P., *et al.*, 2019. Modeling population density using a new index derived from multi-sensor image data. *Remote Sensing*, 11 (22), 2620.
- Luo, P., *et al.*, 2022. Identifying determinants of spatio-temporal disparities in soil moisture of the northern hemisphere using a geographically optimal zones-based heterogeneity model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 185, 111–128.
- Luo, P., Song, Y., and Wu, P., 2021. Spatial disparities in trade-offs: economic and environmental impacts of road infrastructure on continental level. *GIScience & Remote Sensing*, 58 (5), 756–775.
- Ma, T., *et al.*, 2015. Night-time light derived estimation of spatio-temporal characteristics of urbanization dynamics using DMSP/OLS satellite data. *Remote Sensing of Environment*, 158, 453–464.
- Matheron, G., 1963. Principles of geostatistics. *Economic Geology*, 58 (8), 1246–1266.

- Mitas, L., and Mitsova, H., 1999. Spatial interpolation. In: P. Longley, M.F. Goodchild, D.J. Maguire, and D.W. Rhind, eds. *Geographical information systems: principles, techniques, management and applications*, vol. 1. New York, NY: Wiley, 481–492.
- Oliver, M.A., and Webster, R., 1990. Kriging: a method of interpolation for geographical information systems. *International Journal of Geographical Information Systems*, 4 (3), 313–332.
- Preston, R.E., 1966. The zone in transition: a study of urban land use patterns. *Economic Geography*, 42 (3), 236–260.
- Sadahiro, Y., 2000. Accuracy of count data transferred through the areal weighting interpolation method. *International Journal of Geographical Information Science*, 14 (1), 25–50.
- Song, Y., 2022. The second dimension of spatial association. *International Journal of Applied Earth Observation and Geoinformation*, 111, 102834.
- Song, Y., et al., 2018. Segment-based spatial analysis for assessing road infrastructure performance using monitoring observations and remote sensing data. *Remote Sensing*, 10 (11), 1696.
- Song, Y., et al., 2020a. An optimal parameters-based geographical detector model enhances geographic characteristics of explanatory variables for spatial heterogeneity analysis: cases with different types of spatial data. *GIScience & Remote Sensing*, 57 (5), 593–610.
- Song, Y., et al., 2020b. A spatial heterogeneity-based segmentation model for analyzing road deterioration network data in multi-scale infrastructure systems. *IEEE Transactions on Intelligent Transportation Systems*, 22 (11), 7073–7083.
- Song, Y., and Wu, P., 2021. An interactive detector for spatial associations. *International Journal of Geographical Information Science*, 35 (8), 1676–1701.
- Tobler, W., 2004. On the first law of geography: a reply. *Annals of the Association of American Geographers*, 94 (2), 304–310.
- Wang, J.F., et al., 2010. Geographical detectors-based health risk assessment and its application in the neural tube defects study of the Heshun Region, China. *International Journal of Geographical Information Science*, 24 (1), 107–127.
- Wang, J.F., Zhang, T.L., and Fu, B.J., 2016. A measure of spatial stratified heterogeneity. *Ecological Indicators*, 67, 250–256.
- Wong, T.T., 2015. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, 48 (9), 2839–2846.
- Yoo, E.H., and Kyriakidis, P.C., 2006. Area-to-point kriging with inequality-type data. *Journal of Geographical Systems*, 8 (4), 357–390.
- Zhang, Z., Song, Y., and Wu, P., 2022. Robust geographical detector. *International Journal of Applied Earth Observation and Geoinformation*, 109, 102782.
- Zhu, D., et al., 2020. Spatial interpolation using conditional generative adversarial neural networks. *International Journal of Geographical Information Science*, 34 (4), 735–758.

Appendix A

$$\begin{aligned}
 E(\hat{Z}_0 - Z_0) &= E\left(\sum_{i=1}^n \lambda_i Z_i + \sum_{j=n+1}^{n+k} \lambda_j Z_j - Z_0\right) \\
 &= \sum_{i=1}^n \lambda_i E(Z_i) + \sum_{j=n+1}^{n+k} \lambda_j E(Z_j) - E(Z_0) \\
 &= \left(\sum_{i=1}^n \lambda_i - 1\right) * m_{s1} + \left(\sum_{j=n+1}^{n+k} \lambda_j * m_{s2}\right)
 \end{aligned} \tag{A1}$$

where m_{s1} , m_{s2} are the expectations of the variable inside and outside the interpolation stratum, respectively. In this study, m_{s1} is the mean value of observations in the interpolated stratum, and m_{s2} is the mean value of observations outside the interpolated stratum.

Appendix B

The estimation error is transformed into the following equation using the residues:

$$\hat{Z}_0 - Z_0 = \sum_{i=1}^n \lambda_i (Z_i - m_{s1}) + \sum_{j=n+1}^{n+k} \lambda_j (Z_j - m_{s2}) - (Z_0 - m_{s1}) = \sum_{i=1}^n \lambda_i R_i + \sum_{j=n+1}^{n+k} \lambda_j R_j - R_0 \quad (B1)$$

where R_i , R_j , and R_0 are the residues of Z_i , Z_j , and Z_0 after removing the expectations, respectively.

$$\begin{aligned} \delta_E^2 &= \text{var}(\hat{Z}_0 - Z_0) = \text{var} \left[\left(\sum_{i=1}^n \lambda_i R_i + \sum_{j=n+1}^{n+k} \lambda_j R_j \right) - R_0 \right] \\ &= \text{var} \left(\sum_{i=1}^n \lambda_i R_i + \sum_{j=n+1}^{n+k} \lambda_j R_j \right) - 2 \text{Cov} \left[\left(\sum_{i=1}^n \lambda_i R_i + \sum_{j=n+1}^{n+k} \lambda_j R_j \right), R_0 \right] \\ &\quad + \text{var}(R_0) \\ &= \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \text{Cov}_{inside}(R_i, R_j) + \sum_{i=n+1}^{n+k} \sum_{j=n+1}^{n+k} \lambda_i \lambda_j \text{Cov}_{outside}(R_i, R_j) \\ &\quad + 2 \sum_{i=1}^n \sum_{j=n+1}^{n+k} \lambda_i \lambda_j \text{Cov}_{between}(R_i, R_j) - 2 \sum_{i=1}^n \lambda_i \text{Cov}_{inside}(R_i, R_0) \\ &\quad - 2 \sum_{j=n+1}^{n+k} \lambda_j \text{Cov}_{between}(R_j, R_0) + \text{var}(R_0) \end{aligned} \quad (B2)$$

where Cov_{inside} is the spatial covariance function of the stratum to be interpolated, $\text{Cov}_{outside}$ is the spatial covariance function of all strata outside the stratum to be interpolated, and $\text{Cov}_{between}$ is the spatial covariance function between the interpolated stratum and other strata.

To minimize the δ_E^2 as well as achieve the unbiased estimation (Equations (3) and (A1)), the Lagrange multiplier was introduced to solve this optimization problem. The built Lagrange function is as follows:

$$J = \delta_E^2 + 2L \left[\left(\sum_{i=1}^n \lambda_i - 1 \right) * m_{s1} + \left(\sum_{j=n+1}^{n+k} \lambda_j * m_{s2} \right) \right] \quad (B3)$$

There are three unknown variable sets: λ_i ($i = 1, 2, \dots, n$), λ_i ($i = n + 1, n + 2, \dots, n + k$), and L . The matrix includes totally $n + k + 1$ unknown values. Solving the matrix using the Lagrange multiplier as follows:

$$\begin{cases} \frac{\partial J}{2\partial \lambda_i} = \sum_{j=1}^n \lambda_j \text{Cov}(R_i, R_j) + \sum_{j=n+1}^{n+k} \lambda_j \text{Cov}(R_i, R_j) + L = \text{Cov}(R_0, R_i) (i = 1, 2, \dots, n) \\ \frac{\partial J}{2\partial \lambda_i} = \sum_{j=n+1}^{n+k} \lambda_j \text{Cov}(R_i, R_j) + \sum_{j=1}^n \lambda_j \text{Cov}(R_i, R_j) + L = \text{Cov}(R_0, R_i) (i = n + 1, n + 2, \dots, n + k) \\ \left(\sum_{i=1}^n \lambda_i - 1 \right) * m_{s1} + \left(\sum_{j=n+1}^{n+k} \lambda_j * m_{s2} \right) = 0 \end{cases} \quad (B4)$$

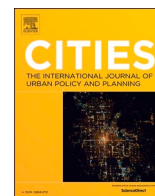
These three equations are transformed as the matrix to be clearly understood, as follows:

$$\begin{bmatrix} R_{1,1} & \dots & R_{1,n+k} & m_{s1} \\ R_{2,1} & \dots & R_{2,n+k} & m_{s1} \\ \dots & \dots & \dots & \dots \\ R_{n+k,1} & \dots & R_{n+k,n+k} & m_{s2} \\ m_{s1} & \dots & m_{s2} & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \dots \\ \lambda_{n+k} \\ L \end{bmatrix} = \begin{bmatrix} R_{1,0} \\ R_{2,0} \\ \dots \\ R_{n+k,0} \\ m_{s1} \end{bmatrix} \quad (B5)$$

where $R_{i,j}$ represents the $\text{Cov}(R_i, R_j)$.

A7. Extracting the crime information implied in the built environment by treating it as the anomalies

Reference: Yao, Y., Dong, A., Liu, Z., Jiang, Y., Guo, Z., Cheng, J., Guan, Q. and Luo, P.*, 2023. Extracting the pickpocketing information implied in the built environment by treating it as the anomalies. *Cities*, 143, p.104575.



Extracting the pickpocketing information implied in the built environment by treating it as the anomalies

Yao Yao^{a,b}, Anning Dong^a, Zhiqian Liu^a, Ying Jiang^a, Zijin Guo^a, Junyi Cheng^c, Qingfeng Guan^a, Peng Luo^{d,*}

^a School of Geography and Information Engineering, China University of Geosciences, Wuhan 430078, Hubei Province, China

^b Center for Spatial Information Science, The University of Tokyo, Kashiwa-shi, Chiba 277-8568, Japan

^c Institute of Remote Sensing and Geographic Information Systems, Peking University, 5 Summer Palace Road, Beijing 100871, China

^d Chair of Cartography and Visual Analytics, Technical University of Munich, 80333 Munich, Germany

ARTICLE INFO

Keywords:

Pickpocketing crime
Street view
Deep anomaly detection
Interpretable analysis
Social disorder

ABSTRACT

The practice of crime risk mapping, enabled by the utilization of geospatial big data such as street view images, has received significant research attention. However, in situations where available data is scarce, mapping models may suffer from underfitting and generate inaccurate spatial pattern estimations of crime risk. The covert nature of pickpocketing crimes results in limited observed areas relevant to such criminal events, leading to insufficient coverage of geospatial data. Moreover, the location of crime is also influenced by socio-economic characteristics that may introduce biases into crime risk estimates. These factors render it challenging for the model to capture a valid crime risk pattern, potentially yielding misleading conclusions. Therefore, effectively extracting crime risk with limited data remains a challenge, especially when relying on easily accessible, widespread, and unbiased geospatial data. To address this challenge, we propose a novel crime risk assessment framework based on deep anomaly detection techniques, assuming that urban landscape anomalies carry deep crime risk information. We take Shenzhen as the study area and map the distribution of pickpocketing risk using street view images, accurately revealing the spatial aggregation of pickpocketing crime risk. Our findings indicate that pickpocketing crime in China is caused by regional economic conditions, built environment factors, and human routine activities. This study provides valuable insights for policing and prevention strategies aimed at addressing pickpocketing crimes in large Chinese cities. By leveraging our proposed crime risk assessment framework, decision-makers can allocate resources more efficiently and develop targeted interventions to mitigate crime risks.

1. Introduction

Crime has a significant impact on economic growth and human lives (ToppiReddy et al., 2018), a problem that has long plagued human societies. One of the most common crimes is Pickpocketing, which involves stealing a victim's property in a public or semi-public place (Deshotels, 2013). Pickpocketing is characterized by high concealment, small amounts of money involved, and high significant financial and material resources investment in detection and apprehension (Lafree & Birkbeck, 2010). Therefore, preventing pickpocketing yields greater policing benefits than detecting and apprehending the offender. The social disorder theory and crime pattern theory suggest out that an

objective environment can stimulate crime generation to some extent (Shaw et al., 1942). Since pickpocketing requires physical contact with the victim and often occurs in urban environments, exploring the association between the urban environment and pickpocketing is vital for police departments to prevent and control this type of crime and maintain social stability.

Several previous studies have assessed crime risk using historical case data, spatio-temporal environmental data, or behavioral trajectory data. However, these approaches present certain limitations. Historical case data-based assessments can only consider crime patterns based on actual case occurrences and have low dimensions and a single source of information, making it challenging to assess crime risks in areas where

* Corresponding author.

E-mail addresses: yaoy@cug.edu.cn (Y. Yao), donganning@cug.edu.cn (A. Dong), liuzhiqian@cug.edu.cn (Z. Liu), snotra@cug.edu.cn (Y. Jiang), gzj2017@cug.edu.cn (Z. Guo), junyicheng@pku.edu.cn (J. Cheng), guanqf@cug.edu.cn (Q. Guan), peng.luo@tum.de (P. Luo).

<https://doi.org/10.1016/j.cities.2023.104575>

Received 20 May 2022; Received in revised form 18 August 2023; Accepted 21 September 2023

0264-2751/© 2023 Elsevier Ltd. All rights reserved.

no crime has occurred or where crime data are unavailable (Hossain et al., 2020; Hu et al., 2018). Some researchers have integrated spatio-temporal environmental data (Ding & Zhai, 2021; Giménez-Santana et al., 2018) to further consider the spatio-temporal effects of the background environment on crime generation and evolution. Others have used micro-level behavioral trajectory data (Rumi et al., 2019; Xiao & Zhou, 2020) to assess crime risk, incorporating socioeconomic data such as demographic, GDP, and unemployment rates and trajectory data such like location check-in and cab traffic for crime risk assessment. However, obtaining fine-scale residential travel and socioeconomic data can be challenging in some areas, and, some environmental data may also be difficult to collect, leading to limited applicability and generalizability of existing methods. Thus, addressing the question of how to use easily accessible and equally objective indicators that can be correlated and mapped to crime instead of hard-to-obtain economic indicators remains a challenge.

In the field of crime risk mapping, street view imagery offers a potential solution due to its extensive coverage. Recent studies have shown that street view images can provide insight into the physical urban environment and reveal crime risk (He et al., 2017; Zhang et al., 2021). Street view imagery accurately depicts the physical urban environment and allows for inferences about urban perception (Wang et al., 2019a; Yao et al., 2019). With easy availability, high-frequency updates, and microscopic perspectives on the city, street view imagery has become an increasingly popular tool for analyzing human or physical environmental elements using semantic segmentation or target recognition methods to test crime theories (He et al., 2017; Yue et al., 2022). Zhang et al. (2021) recently analyzed Houston street view imagery and historical criminal records and found a discrepancy between people's perception of safety in the urban environment and the actual crime rate. However, most previous studies require large amounts of real, tagged crime data for analysis, which can be challenging to obtain for sparsely located crime events such as pickpocketing, whose data may have biased spatial distribution. Therefore, it remains unclear whether the relationship between street view imagery and crime can be effectively mined when dealing with sparse and biased data.

Pickpocketing is a very typical and common type of crime that affects people's daily lives, yet reliable data regarding its occurrence is scarce and biased. The available data may not accurately reveal the true spatial patterns of crime risk due to several factors. Firstly, crime data is scarce and incomplete in certain regions. For instance, in China, publicly available crime data primarily consists of judgment documents, which do not always reflect the true number of pickpocketing crimes committed. Due to the relatively minor nature of pickpocketing offenses, suspects often employ various methods to evade surveillance and avoid detection, resulting in underreporting of such crimes. Secondly, there is a problem of biased sampling in the available data. The spatial distribution of crime locations in judgment documents is not solely determined by the risk of crime but can also be influenced by population density, law enforcement efforts, economic conditions, and other attributes. As an example, densely populated and more economically developed areas have a high density of crime points, while sparsely populated suburban or rural areas may have limited data on pickpocketing, despite not necessarily having lower crime risks. Moreover, obtaining a conviction for a pickpocketing offense involves a complex process that includes the occurrence of the crime, police investigation, and court proceedings. Therefore, although judgment documents can serve as a reference for analyzing crime patterns, they may not fully capture the real spatial pattern of pickpocketing crime.

In conclusion, we contend that publicly available crime data, especially for pickpocketing, does not provide a comprehensive representation of the true spatial pattern of crime risk. Firstly, such data is highly sparse in space, which limits its utility in producing accurate crime risk assessments. Secondly, the pattern of crime data suffers from sampling bias, and can be influenced by socio-economic factors beyond crime risk considerations. While many studies have utilized multiple data sources

to analyze crime risk, these studies often require large quantities of labeled data for training models (Hajela et al., 2021; Xiao & Zhou, 2020). However, given the limited amount of labeled crime data and the significant bias present at crime points, there are currently few effective methods for achieving accurate spatial predictions of global crime risk. Despite studies indicating that the collection of crime information by law enforcement agencies inevitably suffers from biases due to influences from the agencies themselves and those reporting the crimes, it is important to note that these data sources still exhibit fewer random biases compared to other sources, such as spontaneously reported crime victim survey data. Moreover, they provide accurate records of crime locations and processes, thus remaining a more trustworthy source of crime data (Brunton-Smith et al., 2023; Buil-Gil et al., 2022). However, when crime data is accurate but scarce, it remains unclear to what degree policing levels, as quantified by crime data such as judgment documents, can be trusted as reliable indicators of crime risk. Given the current limitations associated with relying solely on real crime points for analysis and decision support, we propose the following research question: How can precise predictions of global crime risk be generated when crime data are sparsely sampled and biased? If it proves possible to accurately extract crime risk information from such sparse data sources, particularly in relation to hidden crimes like pickpocketing, this could have significant implications for large-scale crime risk assessment and urban governance.

Due to various factors, it is common to conduct research on data with bias in the field of geographic information. Whether it's bias brought about in the data collection process (Li et al., 2016; Zhang & Zhu, 2019a), bias in geographically large data voluntarily uploaded by the public (Zhang, 2022; Zhang & Zhu, 2019b), or even bias in data collected by government agencies (Brunton-Smith et al., 2023; Buil-Gil et al., 2022), there are inevitable deviations. Although the data may be geographically biased, it is still numerically correct. We can trust that the more similar the geographical configuration (i.e., spatial neighborhood geographical variables) of two points (regions), the more similar the values (processes) of the target variable at these two points will be (Zhu et al., 2018). Based on this idea, finding suitable environmental features for the data and designing analysis methods that adapt to this data has become the key task in using biased data for geographical modeling.

This study proposes the Crime Anomaly Detection based on Street View (CADSV) framework, which utilizes deep learning methods to tackle the aforementioned challenges. Anomaly detection is a popular technique for identifying rare or unusual patterns within large datasets (Chandola et al., 2009), which is similar to a crime assessment task that extracts risk information from limited crime labeled street view images. In this study, we focus on the city of Shenzhen where we assess the risk of pickpocketing at various locations using judgment documents as to the supporting data. To further investigate the socioeconomic factors associated with crime, we incorporate point of interest (POI) data is used to represent the urban functional structures. Additionally, the random forest and SHapley Additive exPlanations (SHAP) techniques are used to utilize the complex relationship between the urban socioeconomic structure and spatial environment.

2. Related work

2.1. Risk assessment of pickpocketing crime based on spatial analysis

Crime risk assessment is essential in policing (Fan et al., 2021; Oswald et al., 2018). Some scholars have conducted crime risk assessments based on analysis of historical case data (Hossain et al., 2020; Hu et al., 2018). For instance, Hu et al. (2018) utilized a spatiotemporal kernel density estimation (STKDE) method to analyze the history of crimes in a particular location and identify burglary hotspots in the region. Similarly, Hossain et al. (2020) employed decision trees and the k-nearest neighbors (KNN) algorithm to evaluate crime risk using San

Francisco's criminal activity data from 2003 to 2015. However, these studies do not account for the interaction between crime and other social environment factors. The data used in these studies are typically low-dimensional and obtained from single sources, rendering them suitable only for macro trend statistical analyses with limited explanatory power for crime risk assessment. Furthermore, reliance on historical crime data from a specific location may limit the transferability of findings beyond that context.

The Broken Window Theory (BWT) elucidates the relationship between crime and environment, positing that visible signs of disorder and neglect can foster further criminal activity, including serious crimes (Wilson & Kelling, 1982). Certain scholars have augmented historical case data with spatiotemporal environmental data to better consider the spatiotemporal effects of the contextual environment on the generation and evolution of crime (Ding & Zhai, 2021; Giménez-Santana et al., 2018). For example, Giménez-Santana et al. (2018) used a risk-topography modeling approach to identify environmental factors associated with three types of violent crime events (homicide, assault, and theft) and assessed the risk for different crime types. Ding and Zhai (2021) used crime statistics and observed climate records in Beijing to demonstrate strong correlations between PM2.5, the Air Quality Index (AQI), and bus pickpocketing crimes. Based on these findings, they utilized a support vector machine approach was used to predict the risk of bus pickpocketing crimes. Many studies have demonstrated that crime generally tends to concentrate in micro-specific locations such as streets, thereby highlighting the importance of assessing crime risk at the micro-level for effective crime prevention and police control (Groff et al., 2010; Weisburd et al., 2004). However, spatio-temporal environmental data are often collected at the grid scale, which has limited spatial resolution, and is generally only suitable for macro-level studies while being insufficiently assessed at the micro-scale.

Crime pattern theory suggests that offenders do not randomly search for potential targets but instead rely on the path or routes of their daily activities to find suitable targets (Bernasco et al., 2013; Bernasco et al., 2017; Brantingham & Brantingham, 2013). Therefore, some scholars have integrated suspects' behavioral trajectory data with spatio-temporal environmental data (Bouma et al., 2014; Rumi et al., 2019; Xiao & Zhou, 2020). Notably, Zhao and Tang (2017) employed POI check-in data, weather data, and public service complaint data to predict future crimes in New York City. Results showed that the inclusion of dynamic data characterizing daily human activity helped to accurately assess crime risks. Hajela et al. (2021), meanwhile, constructed distinct crime prediction models using taxi data, historical crime data, and demographic data, comparing their effectiveness against each other. This study demonstrated that methods incorporating dynamic data are more effective in crime prediction than those relying exclusively on crime data or data pertaining to social environmental factors. In summary, these studies consider the impact of the environment on crime at a finer scale, which is more effective in assessing pickpocketing risk at the micro-scale. However, such research typically requires low-accessibility data, thereby limiting its applicability to larger areas. Conversely, street view data can satisfy both environmental information provision and large-scale information provision, providing data support for crime risk assessment.

2.2. Street view image and city perception

Street View Images (SVI) are composed of panoramic images of various locations on the street that provide a comprehensive reflection of the physical urban environment and human activities on a large scale (Kang et al., 2020; Yao et al., 2019). In comparison to behavioral trajectory data, SVIs are low-cost and highly accessible. Additionally, they can capture detailed information in the physical environment more comprehensively using a perspective similar to that of the human eye (Zhang et al., 2020). As such, they have been integrated into diverse urban studies, including urban safety perceptions (Wang et al., 2019b;

Zhang et al., 2021) and urban crime research (He et al., 2017). For instance, He et al. (2017) employed used Google Street View to identify factors in the physical environment of Columbus cities that contribute to violent criminal activity. Results showed positive associations between crime rates and street graffiti, abandoned buildings, and abandoned cars. Similarly, Zhang et al. (2021) analyzed street view images and historical criminal records in Houston, finding that areas where people feel unsafe do not correlate with high crime rates. There existed a perceived bias between perceived safety and actual crime rates in the urban environment.

Most of the current studies examining the relationship between street view images and crime risk have focused on Western cities. However, it remains uncertain questionable whether research findings on Western cities can be effectively applied to Chinese cities. Firstly, there are significant disparities in architectural and urban planning styles between the East and West (Ashihara & Riggs, 1983). Secondly, various socio-political factors contribute to the differences in crime patterns between East and West (Farrell & Bouloukos, 2001; Steffensmeier et al., 2017). For instance, a comparative study of high school students in China and the United States revealed that crime rates were significantly lower in Chinese schools were much lower than in American ones (Webb et al., 2011). As such, it is crucial to investigate the relationship between urban environments and crime risk in China using street view images.

Regarding the use of street views for crime risk prediction, a typical approach involves first extracting high-dimensional semantic features from the images, followed by constructing regression models to establish the relationship between these features and crime risk. For example, semantic segmentation can be used to extract the proportion of green space within an image, while target detection can estimate the number of people present (Hipp et al., 2021; Jing et al., 2021). Street view images contain vast amounts of semantic information that humans have yet to explicitly express. This implicit information has the potential to uncover crime risk. However, current solely rely on human-defined semantic features, thus overlooking a large amount of semantic information present in the images. Moreover, the above framework encounters difficulties when analyzing risks in areas not covered by LBS data, particularly with biased and spatially sparse crime data. To address this problem, it is crucial to extract crime risks from street views based on sparse data, which would fill this gap and facilitate the building of end-to-end models by eliminating complex image processing steps.

2.3. Geographical research based on biased data

In fact, in the field of geographic information, analyzing data with a geographical distribution bias is a widely studied problem. For example, Volunteer Geographic Information (VGI), voluntarily uploaded by citizens, is one of the newly emerged types of big geographic data in recent years (Zhang & Zhu, 2018). With its rich geographic information, high update frequency, and low cost, VGI is used to reveal spatiotemporal patterns of geographic phenomena. However, since the spatial distribution of volunteer observation work is neither random nor regular, the observation results are often spatially biased towards areas with high population density or high route accessibility, leading to bias in geographic distribution. To address this issue, researchers typically start by comparing sample locations with the environmental covariates of the predicted areas, improving sample representativeness by comparing similarities. Based on this concept, studies by (Zhang & Zhu, 2019b) and (Zhang, 2022) have mitigated the spatial bias in samples based on VGI by improving sample representativeness.

Additionally, studies based on geographically biased data in the field of digital soil mapping underscore the importance of environmental covariates. Since the spatial distribution of soil samples may lean towards specific geographical areas and be influenced by the personnel taking measurements, soil samples are a type of data that can be easily affected by spatial bias (Li et al., 2016; Zhang & Zhu, 2019a). To solve this problem, (Fan et al., 2020) proposed the SoLIM-FilterNA method,

which predicts the soil property values of unknown units by learning the characteristics of one or more environmental covariates, as long as the uncertainty of that unit does not exceed the set threshold. (Zhu et al., 2018) proposed a method that does not use the explicit relationships derived from the entire sample set, but instead makes predictions based on the comparison of the geographical configurations of the sample points and the prediction points. This study suggests that accurate spatial predictions can be made based on biased samples. Similarly, facing the issue of data sparsity, (Du et al., 2020) proposed a semi-supervised machine learning method for predictive mapping, which uses the natural aggregation (clustering) pattern of environmental covariate data to supplement the limited samples in prediction. The characteristic of these studies is that they step outside of the spatial dimension to deal with spatially biased data. It can be seen that in the case of geographical bias, if suitable environmental features and methods can be selected to fit the data, results better than traditional geographic models can be achieved.

2.4. Deep anomaly detection model

Anomaly detection models are commonly used to identify events that have a low probability of occurrence but often cause fatal harm to the system (Chandola et al., 2009). Since crimes tend to be concentrated in specific areas, only a few street images are spatially associated with crime events (Weisburd, 2015). Therefore, the crime risk information contained within the street view images can be classified as anomaly. Anomaly detection tasks that distinguish between anomalous and normal data is a One-Class Classification (OCC) tasks. Early OCC research focused on using statistical methods for feature extraction and developing classifiers. Since 2017, deep learning methods have become the mainstream of OCC research (Perera et al., 2021) which have made progress in several areas, such as cybersecurity intrusion detection (Kim & Kim, 2021), medical pathology image detection (Schlegl et al., 2017), One approach to deep learning-based OCC is to learn normal features and compare differences between test data and normal features, with greater variations indicating anomalous data (Ruff et al., 2018).

The convolutional neural network (CNN) structures can be utilized to achieve high accuracy in anomaly detection algorithms for images (Minhas & Zelek, 2019). Cohen and Hoshen (2020) utilized pyramidal neural networks to detect anomalous images and localize anomalous parts. Massoli et al. (2021) proposed the MOCCA framework to extract features at different depths of deep neural networks, thereby enhancing network discrimination in single-classification (OCC) problems. Sabokrou et al. (2018) introduced the first single-classification model based on GAN networks, which enhances the interpreter’s normalization ability while iteratively reconstructing features.

3. Materials and method

Fig. 1 depicts a flowchart illustrating the pickpocketing crime risk assessment with coupled street view images using deep anomaly detection. The methodology comprises three fundamental stages: (1) Data preparation. Collected the 2018 judgment documents using a web crawler and subsequently extracting the crime locations using a natural language processing models, the crime locations were spatially with the street view images; (2) Mapping urban crime risk using the proposed Constructed Crime Anomaly Detection framework based on Street View (CADSV). We evaluated the risk of pickpocketing crimes by calculating image feature similarities between street view image; (3) Model interpretability analysis. Used POI kernel density data to characterize the drivers of the pickpocketing crime risk. This analysis utilized the Random Forest and SHAP models for interpretability. Additionally, we explored whether these drivers are consistent with the objective environmental risks characterized by the Street View imagery.

3.1. Study area and data

Shenzhen (Fig. 2) is a typical migrant city and the most developed city in South China, consisting of 10 districts. There are significant differences in economic development between downtown and suburban areas in Shenzhen (Meyer, 2016). The downtown areas are Shenzhen’s political, economic, and cultural centre, including Nanshan District,

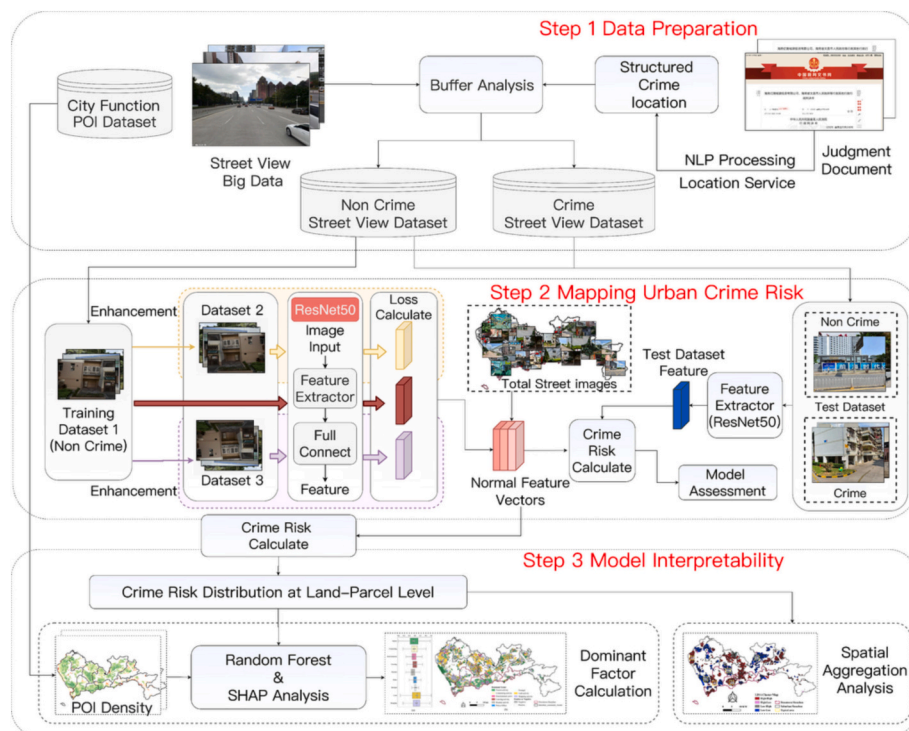


Fig. 1. Schematic overview of pickpocketing crime risk assessment with coupled street view images using deep anomaly detection.

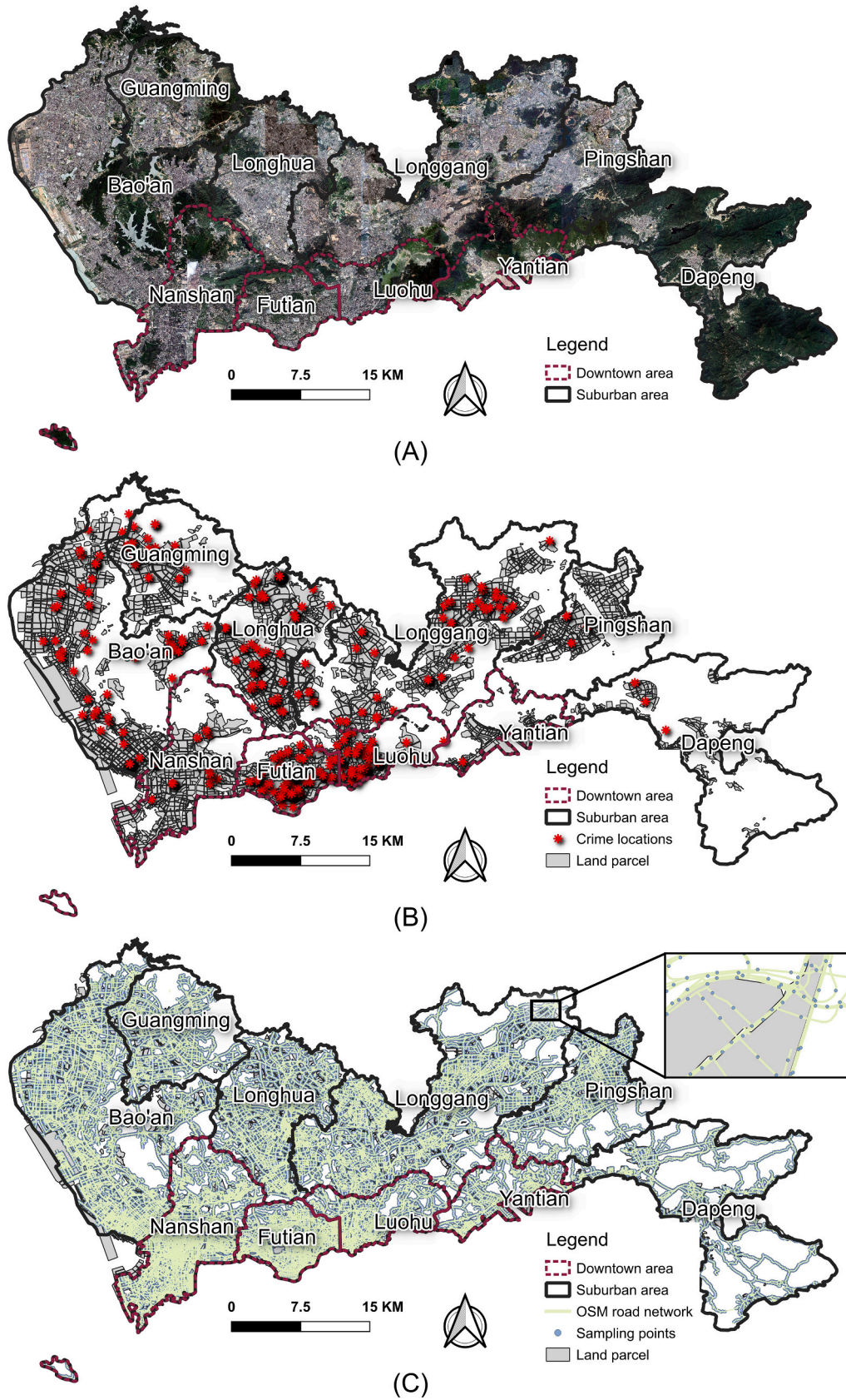


Fig. 2. (A) High-spatial resolution remote sensing imagery and (B) Crime locations and land-parcels in the study area (Shenzhen). (C) Road network and street view image sampling points.

Futian District, Luohu District, and Yantian District. Meanwhile, the suburban areas comprise Longhua District, Longgang District, Pingshan District, Dapeng District, Guangming District, and Baoan District, with a complex composition of foreign and migrant populations. It should be noted that the administrative division of Shenzhen underwent significant reorganization in 2018. To ensure study validity and offer support for future research efforts, all data used in this study were collected from 2018.

Land use planning parcels represent the fundamental unit of urban cadastral management in China. This study employed Shenzhen land use parcel data as the primary analytical unit. These parcels consist of 6913 records retrieved

from the Shenzhen Planning and Natural Resources Bureau website (<https://pnr.sz.gov.cn/>).

Acquiring crime locations from social media platforms has been explored in literature by Hipp et al. (2019). However, this approach may not provide credible crime data and therefore needs to be supported by police or official documents. In this study, the pickpocketing data were obtained from the China Judicial Documents website (<http://wenshu.court.gov.cn>). The Supreme People's Court of China mandates that all Chinese courts to publish judgment documents on the web, including information such as the cause, time, and location of the crime. To validate the accuracy of this data, previous studies have analyzed crime cases from different fields (Cai & Xin, 2019; Miao et al., 2016). Our study captured all criminal cases from 2018, which amounted to 7535 cases. Of these, pickpocketing accounted for 9.05 % (or 682) of all sentencing documents. Natural language processing models were utilized to extract pickpocketing crime locations, which are accurate up to the building level or street level can be spatially matched with street view images.

Street View Image has been utilized in prior studies to reflect the physical environment or residents' perceptions of cities (Helbich et al., 2019; Wang et al., 2019b; Zhang et al., 2018). In this study, Baidu Street View images from 2018 were employed to depict the urban environment in Shenzhen. As one of the largest street view service providers available in China, Baidu Street View covers a vast majority of Chinese cities (Kang et al., 2020). This study used the road network data of Shenzhen city in 2018 were obtained via OpenStreetMap using the OpenStreetMap API. Byun and Kim (2022) have noted that acquiring street views at the street level with a distance of 200 m can effectively reveal the urban environment and its changes. There are also studies on crime that use 200 m as a buffer range for data collection, an interval that is considered to best describe the scale of urban communities (Kadar et al., 2016). Therefore, this study employed a sampling approach that involved the collection of road network data at 200 m intervals, thus obtaining a total of 38,717 sampling points. Subsequently, street view images were acquired from four horizontal directions (0°, 90°, 180°, and 270°) to simulate human visual perception. In total, 154,868 street view images were obtained for all sampling points. These images were then labeled as either pickpocketing risk images or non-pickpocketing risk images based on crime locations. Consistent with the sampling interval, a buffer radius of 200 m was selected to identify street images at risk of pickpocketing. Consequently, a total of 2712 street images were flagged as being at risk of pickpocketing. All other street view images were classified as normal images. It is worth noting that each street view image was labeled only once, although some street view images were located within buffer zones of multiple crime locations.

Point of Interest (POI) data have been demonstrated to effectively reflect the socioeconomic and functional structural characteristics of cities (Yao et al., 2017). In this study, POI data were utilized to analyze the relationship between pickpocketing crimes and urban functions at the micro-scale. The POI data used in this study were derived from Gaode Map (<https://www.amap.com/>), one of China's largest online map providers. A total of 213,476 POI data points from the year 2018 were collected in the study area, which were classified into five major categories: Catering & Entertainment, Education & Health care, Industry, Finance & Insurance, and Other Five major categories (Hu &

Han, 2019). These categories were further divided into nine second-level categories, which are Life Services (39,590, 18.55 %), Transportation (37,536, 17.58 %), Landscape (2515, 1.18 %), Police (2544, 1.19 %), Medical Institutions (20,327, 9.52 %), Restaurants (75,030, 35.15 %), Finance (14,165, 6.64 %), Entertainment (17,107, 8.01 %) and Shopping Malls (4662, 2.18 %). We calculated the density of each type of POI density using kernel density analysis.

3.2. Extracting crime information by treating it as the anomalies

3.2.1. The propose of the assumption and the overall framework

Limited availability of crime data in certain regions makes it difficult to associate street view images with criminal activity. Typically, crimes tend to occur in specific locations according to the crime concentration theory (Weisburd, 2015). To address this issue, we propose that crime information can be viewed as anomalies within urban landscapes. Based on this assumption, we developed a Crime Anomaly Detection based on Street View (CADSV) framework for mining pickpocketing risk information from spatially sparse street view images and performing large-scale risk mapping. The framework is threefold (Fig. 3): 1) The normal feature vectors extraction. 20 % of street view images (totally 29,744 images) labeled with non-crime were randomly selected. The ResNet-50 Network was used to extract the normal feature vectors for all street view image. Normal feature vectors include the feature vector extracted for each image. 2) Verify the effectiveness of the extracted normal feature vectors for revolving crime information. 3) Mapping the crime risk for all street view images in the study area.

In this study, we aimed to evaluate the effectiveness of our proposed Crime Anomaly Detection based on Street View (CADSV) framework. To achieve this objective, we randomly selected 29,744 street view images to extract normal feature vectors using the ResNet-50 Network. Subsequently, we selected 10,148 street images to assess the performance of these extracted normal feature vectors.

It is important to note that we included all crime-labeled images in the test set, resulting in a total of 2712 such images. To ensure accurate assessment of the capability of the extracted features in assessing crime risk, we selected four times as many normal-labeled images as crime-labeled ones. Thus, we randomly selected 7436 images for this purpose. It should be noted that there were no strict guidelines for selecting this number; however, we considered 10,148 images to be sufficient for the evaluation process.

3.2.2. Normal feature extraction and feature adaptation

Self-supervised deep anomaly detection is considered a One-Class Classification (OCC) problem. However, when the amount of data is limited, a trained Convolutional Neural Network (CNN) may not effectively capture the semantic information within image dataset, resulting in suboptimal performance. Recent studies have demonstrated that pre-training can improve the effectiveness of model in deep anomaly detection. The CNN network is trained in a larger dataset to get the original feature vectors, which are then adapted features for use in the target dataset.

Feature adaptation aims to map the data from a different source and target domains into a feature space such that they are as similar as possible to each other in that space. Contrast learning is an excellent and effective self-supervised learning method commonly used for the feature adaptation of pre-trained feature extractors (Khosla et al., 2020; Reiss & Hoshen, 2023). The contrast learning method optimizes the prediction task by extracting a dataset x' of batch size N from the training set and training it against the data-enhanced x' with the loss function shown in Eq. 3.1. Where ϕ denotes the feature extraction module used to compute the feature vectors, which in this study represents the ResNet-50 model with normalization added at the last fully connected layer. This is because this method speeds up the convergence of the model and ensures adaptive normalization of the feature data (Yao et al., 2021a).

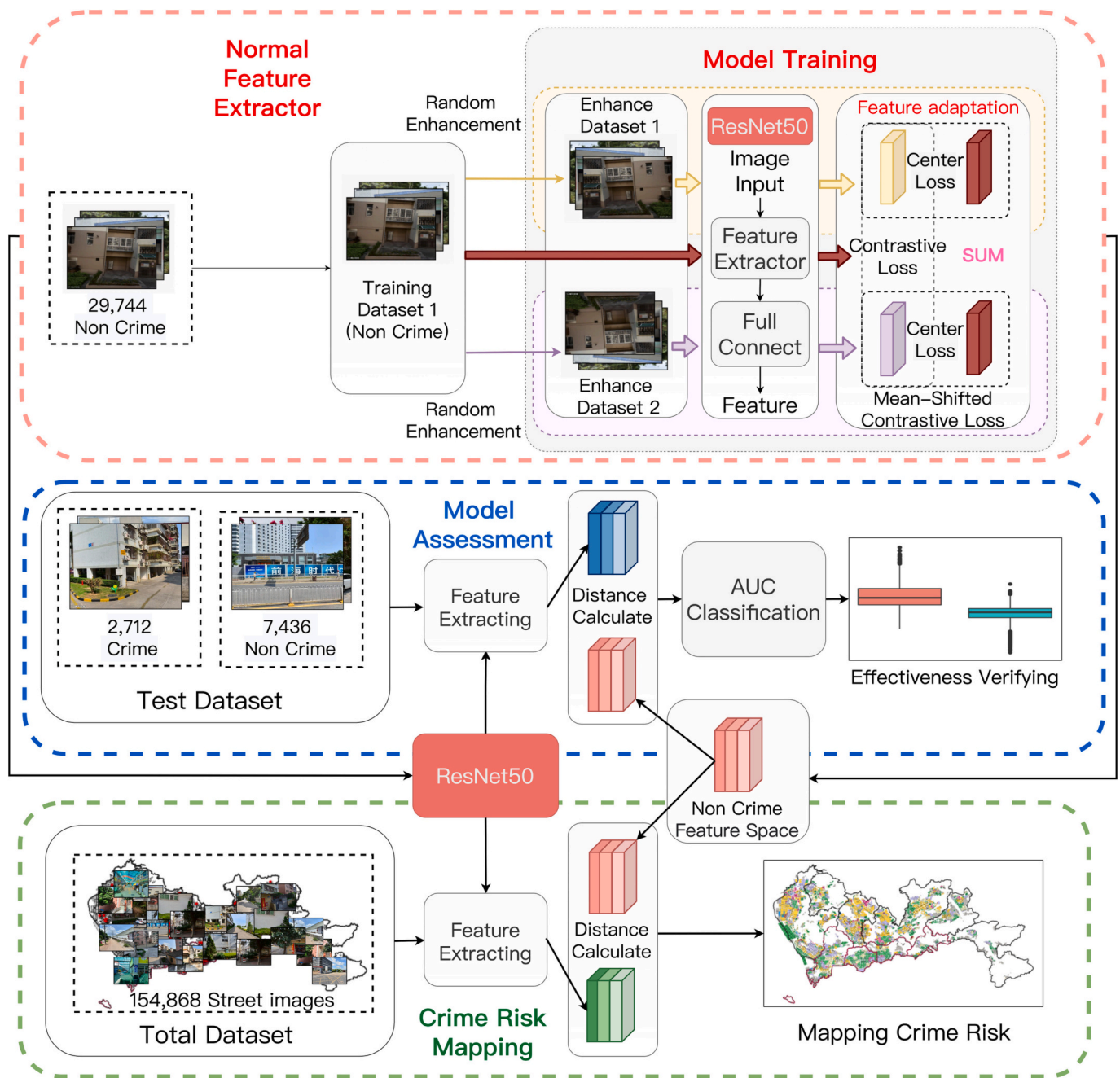


Fig. 3. Crime Anomaly Detection based on Street View (CADSV) framework coupled with Resnet-50 and MSAD.

$$L_{contrastive}(x_i, x_j) = -\log \frac{\exp((\phi(x_i) \cdot \phi(x_j)) / \tau)}{\sum_{i=1}^{2N} \mathbb{1}[x_i \neq x'] \cdot \exp((\phi(x_i) \cdot \phi(x_i)) / \tau)} \quad (3.1)$$

The temperature hyperparameter, which is utilized in contrast learning to regulate the strength of penalty for negative samples (Wang & Liu, 2021), is denoted by τ in Eq. 3.1.

However, although the above contrast learning method is very effective for feature adaptation, for deep anomaly detection in OCC, it may lead to catastrophic collapse, where the accuracy of prediction decreases instead as the number of training increases. Therefore, we used a newly developed loss function, Mean-shifted contrastive loss, proposed by Reiss and Hoshen (2023). It is shown that this solves the problem of dimensional collapse that may occur in the field of image anomaly detection and surpasses the latest previous models in OCC classification. The objective function of mean-shifted loss is shown in Eq. 3.2, and c_{train} denotes the normalized centre of all training images:

$$\theta(x) = \frac{\phi(x) - c_{train}}{\|\phi(x) - c_{train}\|} \quad (3.2)$$

The method is not only able to calculate the Euclidean distance difference between the feature vector of a single image x and c_{train} , but also normalizes the sample difference to the unit sphere and maximizes the distance between negative and positive samples. Besides, in order to reduce the distance between the x' samples and c_{train} after data enhancement, we also introduced the angular center loss (ACL), and the formula is shown in Eq. 3.3:

$$L_{angular}(x) = -\phi(x) \cdot c_{train} \quad (3.3)$$

To sum up, the objective function used in this study that combines the above two constraints is shown in Eq. 3.4:

$$L_{msc}(x', x'') = -\log \frac{\exp((\theta(x') \cdot \theta(x''))/\tau)}{\sum_{i=1}^{2N} \mathbb{1}[x_i \neq x'] \cdot \exp((\theta(x') \cdot \theta(x_i))/\tau)} + L_{angular}(x') + L_{angular}(x'') \quad (3.4)$$

In this study, small batches of data after data enhancement from the original training set are represented by x' and x'' . The data enhancement method includes a series of ways such as flipping, cropping, and Gaussian filtering of the original image features, and ensures that the data enhancement results for x' and x'' are not the same by introducing randomness.

In this study, the ImageNet dataset was selected to pre-train the ResNet-50 network. And the feature adaptation was conducted using the proposed objective function in Eq. (3.4). Each image in the training set was extracted with a feature vector of 2048 dimensions. Normal feature vectors include the feature vectors extracted for each image.

3.2.3. Risk scoring and feature assessment

The process of anomaly scoring (risk scoring) for a street image from the test set is depicted in Fig. 4. Firstly, the pre-trained ResNet-50 network was utilized to extract the feature vector of the test image. Secondly, the KNN model was used to find the K nearest normal vectors for the test feature vector. The K was selected as two, and Euclidean Distance was used as distance metric, referring to the previous work (Reiss et al., 2021). Third, the risk score of the test image was calculated as the cosine distance between the test feature vector and the two nearest normal vectors. The cosine distance can take into account both the Euclidean distance and the angular distance between the features. It has a better deep anomaly detection performance than the other two distance metrics (Reiss et al., 2021). The scoring formula is shown in Eq. 3.5:

$$s(x) = \sum_{\phi(y) \in N_k(x)} 1 - \phi(x) \cdot \phi(y) \quad (3.5)$$

where $N_k(x)$ denotes the k features in the training set that have the closest cosine distance to $\phi(x)$. Moreover, the training set consists of street scenes where no crime occurred, while $s(x)$ indicates the distance of the input street scenes from the normal street scenes. This value ranges between 0 and 1, with higher values indicating a greater probability that the input streetscape belongs to the pickpocketing area.

The Anomaly Scoring was conducted on each image in the test set to assess the ability of the extracted normal feature vectors to characterize non-criminal features. The scoring results of normal-labeled street view images were compared with crime-labeled street view images to verify

the effectiveness of the normal vectors.

The model's performance was evaluated using the AUC metric, which represents the area under the ROC curve (Ling et al., 2003). This metric can address classification result biases towards the majority class when the sample data is imbalanced (Burez & Van den Poel, 2009). The ROC curve is a probability curve that plots the true positive rate (TPR) against the false-positive rate (FPR), while the AUC measures the model's ability to classify correctly. An AUC close to 1 indicates good separability, while an AUC of 0.5 implies no category separation ability.

In this study, the best threshold for classifying whether street view images contain pickpocketing risk features or not was determined using the Youden index (Schisterman et al., 2005; Youden, 1950). Classification accuracy was subsequently assessed using Recall and F1-score. Recall can indicate the proportion of positive samples being correctly predicted in the classification results. On the other hand, the F1-score provides a comprehensive evaluation of classification model accuracy and recall. The formulas used to calculate Recall and F1-score are as follows:

$$Recall = \frac{TP}{TP + FN} \quad (3.6)$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (3.7)$$

where TP denotes the number of correctly identified pickpocketing street view images and FN denotes the number of incorrectly predicted as normal street view images.

3.3. Interpretability analysis based on random forest and Shapley

This study employed the CADSV framework to calculate the street-level pickpocketing crime risk score for each image. The average risk score of street view images within each land parcel was used to characterize the crime risk score in that particular parcel. To investigate the effects of different POIs on the pickpocketing risk, this study utilized a SHAP method to interpretability analysis the results. SHAP (SHapley Additive Explanations) is a data feature analysis method based on game theory (Lundberg & Lee, 2017). This interpretable model that can integrate multiple variables effectively and reveal the contribution of each input spatial data in the model. The SHAP model finds wide application across various fields such as crime (Xie et al., 2022; Zhang et al., 2022) and medicine (Kim & Kim, 2022; Yao et al., 2022). The SHapley values (Štrumbelj & Kononenko, 2014) were calculated in the SHAP model to interpret the contribution and influence of the input

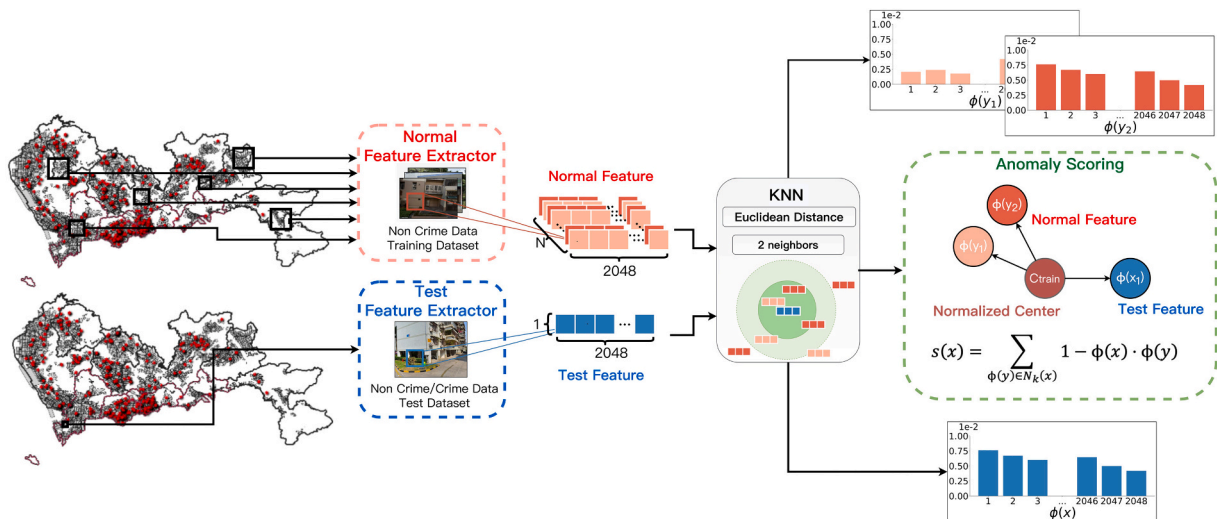


Fig. 4. The process of anomaly scoring for one test image.

features. Specifically, this study employed the Shapley model is used to explain the degree of contribution of different POIs to crime risk. The formula for calculating SHapley values is presented in Eq. 3.8.

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} (v(S \cup \{i\}) - v(S)) \quad (3.8)$$

where ϕ_i denotes the SHapley value of the i -th independent variable and $|N|$ is the number of POI types; S represents an arbitrary disjointly arranged subset of each POI attribute except the i -th variable; $v(S \cup \{i\})$ denotes the output of the model when all data appear; $v(S)$ denotes the output of only the input subset model. Following the above method, the SHapley value of each feature can be calculated by sequentially arranging and sampling each multi-source spatio-temporal data.

In addition to explaining the contribution of spatial variables using the SHAP method, this study used a random forest (RF) model to fit the POI density of different types within each land parcel to the pickpocketing indices. RF model has been used to analyze complex nonlinear correlations between variables in spatial analysis (Hengl et al., 2018; Nussbaum et al., 2018; Rodriguez-Galiano et al., 2012). It can effectively avoid correlation issues in high-dimensional features and has been shown to be the most effective nonlinear fitting model in previous studies (Fernández-Delgado et al., 2014).

4. Result

4.1. Model accuracy

The normal feature vectors were extracted from the training dataset and evaluated in the test dataset. A 5-fold cross-validation was conducted to obtain the hyperparameters of learning rate (0.0005), batch size (64), and epoch (150) were obtained for the training dataset. The risk scores of street view images in the test dataset are shown in Fig. 5. The test dataset include 7436 normal-labeled images and all 2712 crime-labeled images. The results revealed a significant difference in risk scores between crime-labeled and normal-labeled images, with values of 0.41 and 0.29, respectively. Accuracy assessment shows that the AUC, Recall, and F1-Score were 0.921, 0.816, and 0.767, respectively. The scoring result in test dataset demonstrate that the extracted feature vector effectively characterizes the normal urban landscape (Fig. 5) and can detect crime information as anomalies.

Fig. 5c shows the percentage of criminal and non-criminal images in the test set at different intervals of crime risk values. When the crime risk values were less than approximately 0.36, the percentage of non-crime images greatly exceeded that of crime images in each interval of risk values. Conversely, when the crime risk values were >0.36 , the proportion of crime-risk images rapidly increased and surpassed that of

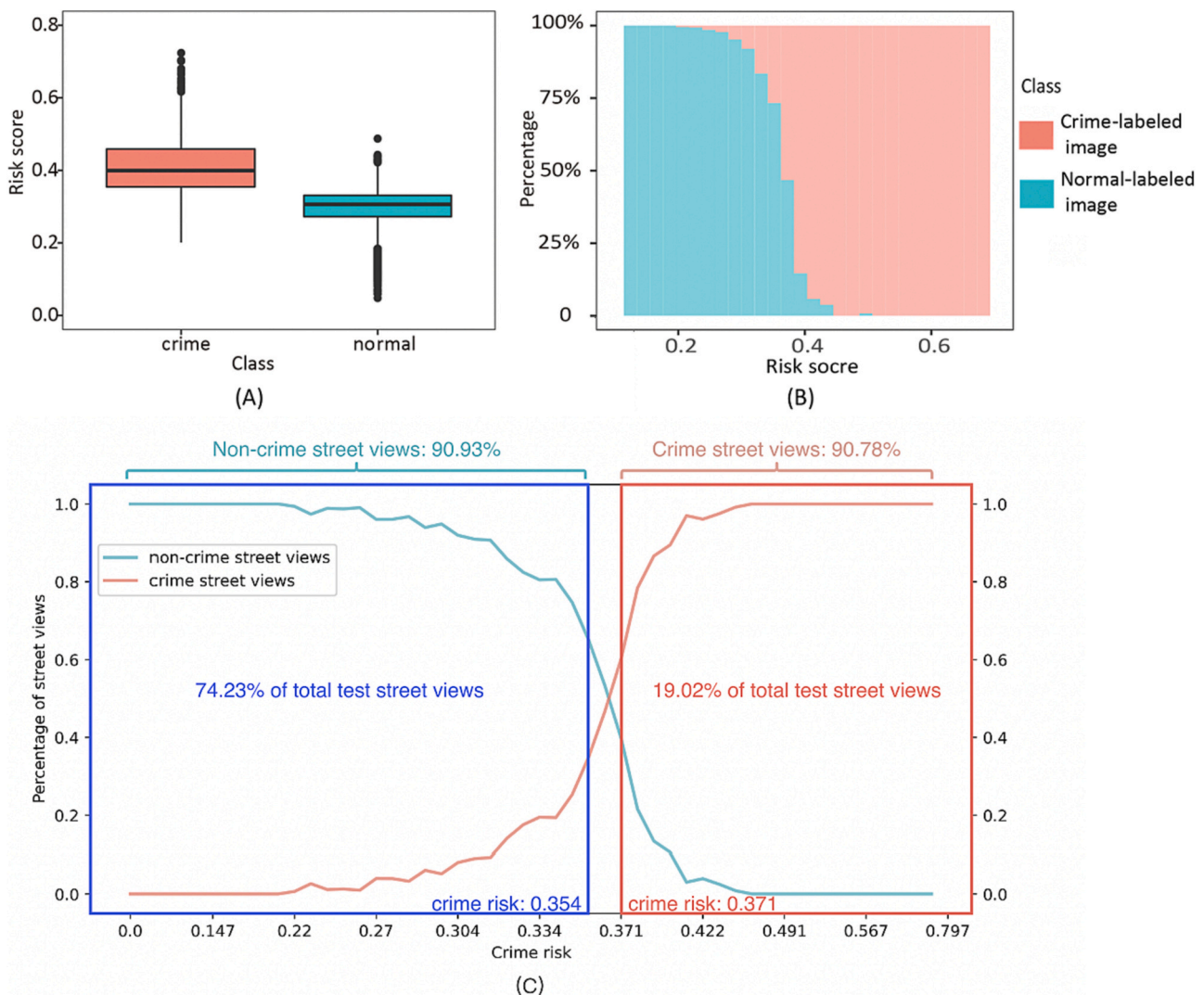


Fig. 5. Risk scores in test datasets: (A) box plot; (B) Percent stacked histogram; (C) Line chart of the percentage of crime/ non-crime street views in each crime risk value interval.

non-crime risk. Therefore, the crime risk values obtained from this model exhibit good discrimination between crime and non-crime images. Further statistical analysis revealed that when the crime risk was <0.354 , including images with a ratio of 74.22 % of cases, 90.05 % of them were non-criminal. In contrast, when the crime risk was >0.371 , 19.04 % of the images were included, while 90.78 % were criminal. At a crime risk value of 0.354–0.371, the ability to distinguish crime images from non-crime images was found to be the weakest, with a ratio of 276:409 between crime and non-crime images in this range. However, this represents only a small percentage (6.75 %) of the total number of images. In conclusion, the crime risk value can distinguish between crime and non-crime images well.

Having established the validity and rationality of the extracted normal feature vectors, pickpocketing crime risk scoring was conducted for all street view images. The resulting scores from street view images were then aggregated into land parcel level. To investigate the driving factors of pickpocketing crime risk, this study used the random forest to fit the nonlinear relationship between each type of POI feature and the pickpocketing crime risk index. In the fitting step, the random forest out-of-bag samples were randomly accounted for 30 %, and the number of decision trees (estimators) was set to 400. The R^2 , RMSE, and MAE were 0.455, 0.016, and 0.868, respectively. These results demonstrated that socioeconomic features revealed by POI data could effectively explain the mapping results of the pickpocketing risk mapping result in most areas.

4.2. Parcel-scale pickpocketing risk mapping

This study mapped the risk distribution of pickpocketing crime for all land parcels in Shenzhen (Fig. 6). The results revealed a high pickpocketing risk in the central city with an average value of 0.362, and a low pickpocketing risk in the peripheral city with an average value of 0.347. This observation suggests that commercial and transportation activities, which are more prevalent in the central urban areas, may play

a significant role in shaping the risk pattern of pickpocketing crime in Shenzhen.

Fig. 7 shows the street view images and their corresponding risk scores of typical functional zones. In general, for each functional area, the more dense the urban building bias, the more chaotic and disorganized the visual perception of the environment, the more likely pickpocketing is to occur. For instance, although the risk of crime inside a factory is low, while a construction site underway is at greater risk. Our findings indicate that the risk of pickpocketing crime is higher in areas relatively disorganized and underdeveloped areas. Tangwei Urban Village, a shantytown in Shenzhen, with a high migrant population and weak security management, had a higher risk of pickpocketing crime (0.404) compared to the resident community Yijing Community (0.354), as seen in Fig. 6. The Business Centre generally had good infrastructure, but its dense flow of people and disorder led to a higher risk of pickpocketing crime. For example, Mixc World Shopping Mall, one of the major business centres in Shenzhen, Mixc World Shopping Mall had a higher risk of pickpocketing crime (0.378) than the average value (0.351). These observations highlight the importance of considering visual perception when evaluating pickpocketing crime risks in urban settings.

The study results highlight variations in pickpocketing crime risk across functional areas. Tourist attractions showed a high average pickpocketing risk value (0.404) due to Shenzhen’s well-developed tourism industry, with numerous scenic green spaces attracting many tourists and providing accessible targets for criminals (Fig. 7). Additionally, the risk values of pickpocketing in residential (0.372), commercial (0.362), industrial (0.368), and school (0.357) areas were higher than the average pickpocketing risk value (0.351). Such functional areas were characterized by dense crowds that offered an opportunity for disorder, making offenders more likely to commit pickpocketing crimes. Parks and landscape spaces were adjacent to residential areas also had many open spaces (Fig. 7 Scenery) that attracted offenders.

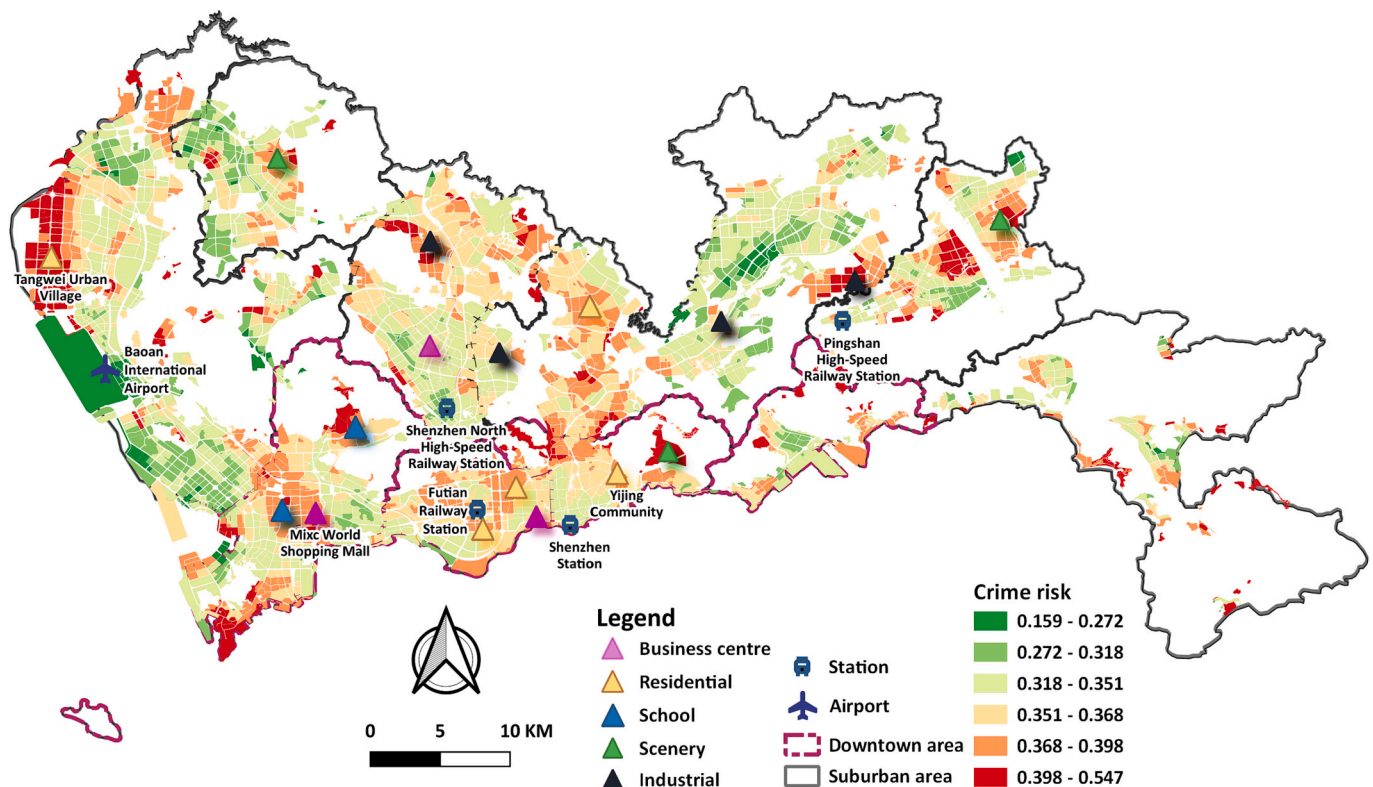


Fig. 6. The distribution of pickpocketing crime risk at land parcel-level in Shenzhen. The triangle marks typical functional areas.



Fig. 7. Street view images of typical functional zones of Shenzhen: The Crime risk axis represents the pickpocketing risk score assessed by the CADSV model, and the Land-use types axis represents typical functional zones.

Moreover, differences in social structure within functional areas can lead to heterogeneity in pickpocketing crime risk. As shown in Fig. 6, the transportation facilities in the city centre had a 39.4 % higher average pickpocketing risk than transportation facilities in other areas. Specifically, transportation facilities such as Shenzhen Station (0.377) and Futian Railway Station (0.373), with an average risk value of 0.375, while those in Baoan International Airport (0.172), Pingshan High-Speed Railway Station (0.331), and Shenzhen North High-Speed Railway Station (0.306), had an average pickpocketing risk level of 0.269. As shown in Fig. 7, the streetscape of stations was very similar in different areas. The risk was higher in the city centre with better economic development, suggesting that economic factors play a complex role in pickpocketing crime.

4.3. Spatial aggregation analysis of pickpocketing crime risk

The global Moran' I index of pickpocketing crime risk in Shenzhen was 0.591 (p -value < 0.001, z -score 51.219), indicating a significant spatial correlation. Furthermore, local spatial autocorrelation analysis based on the Local Moran's I index was conducted to investigate the pattern of urban crime aggregation (Fig. 8). The results indicated that social factors significantly influenced the clustering pattern of pickpocketing crimes. Approximately 29.9 % of areas in Shenzhen had high-high aggregation of a pickpocketing crime risk, which were mainly located in urban central areas, such as University Town (Fig. 8(A)) and Futian CBD (Fig. 8(B)). These regions were characterized by a high concentration of people and wealth, making them prime targets for pickpocketing crimes. Additionally, areas outside the central urban area, such as Tangwei Urban Village (Fig. 8(C)), also showed high-high aggregation due to their inadequate infrastructure construction and a large

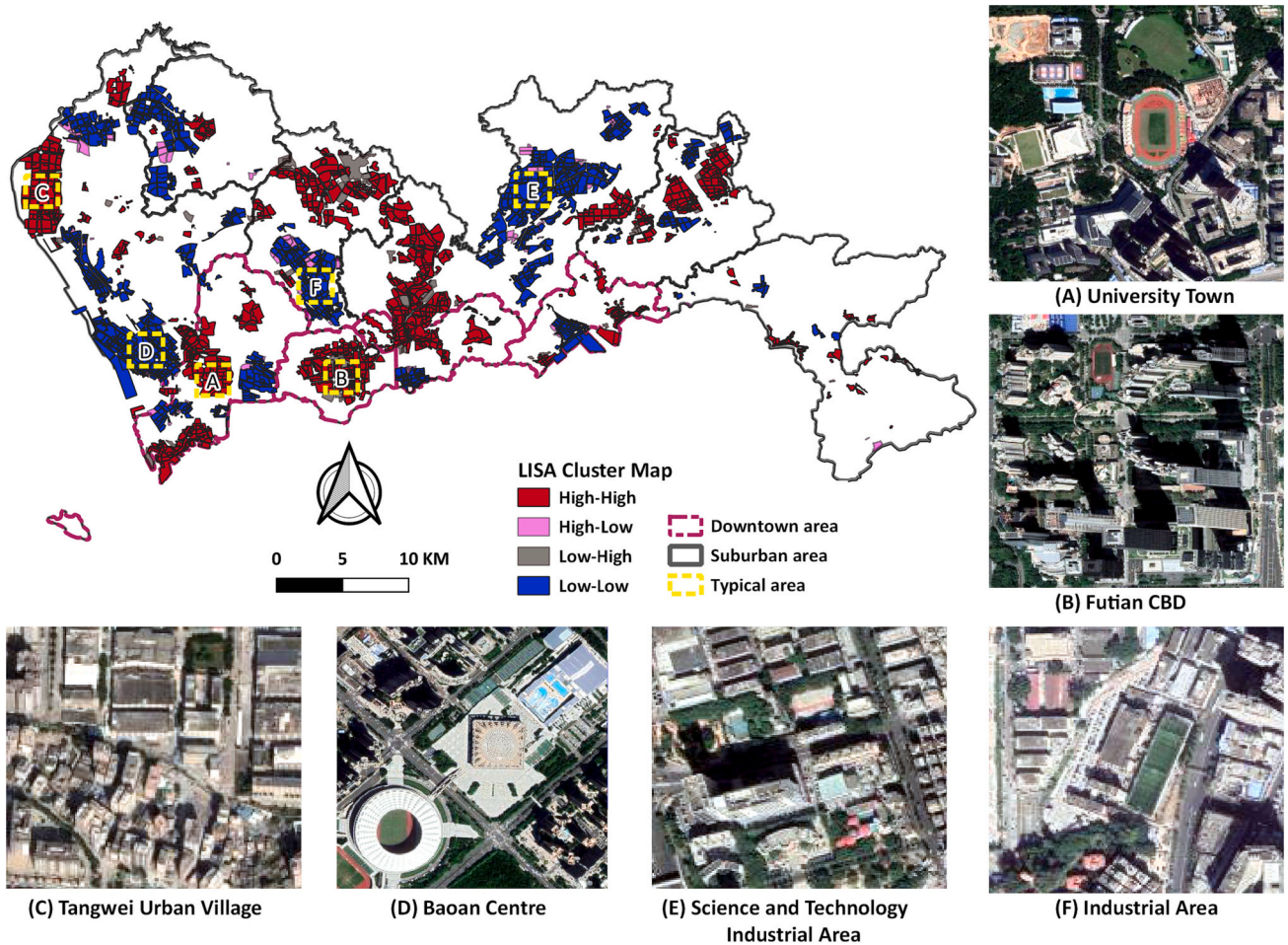


Fig. 8. Results of parcel-scale pickpocketing crime risk aggregation in Shenzhen and remote sensing images of some typical areas.

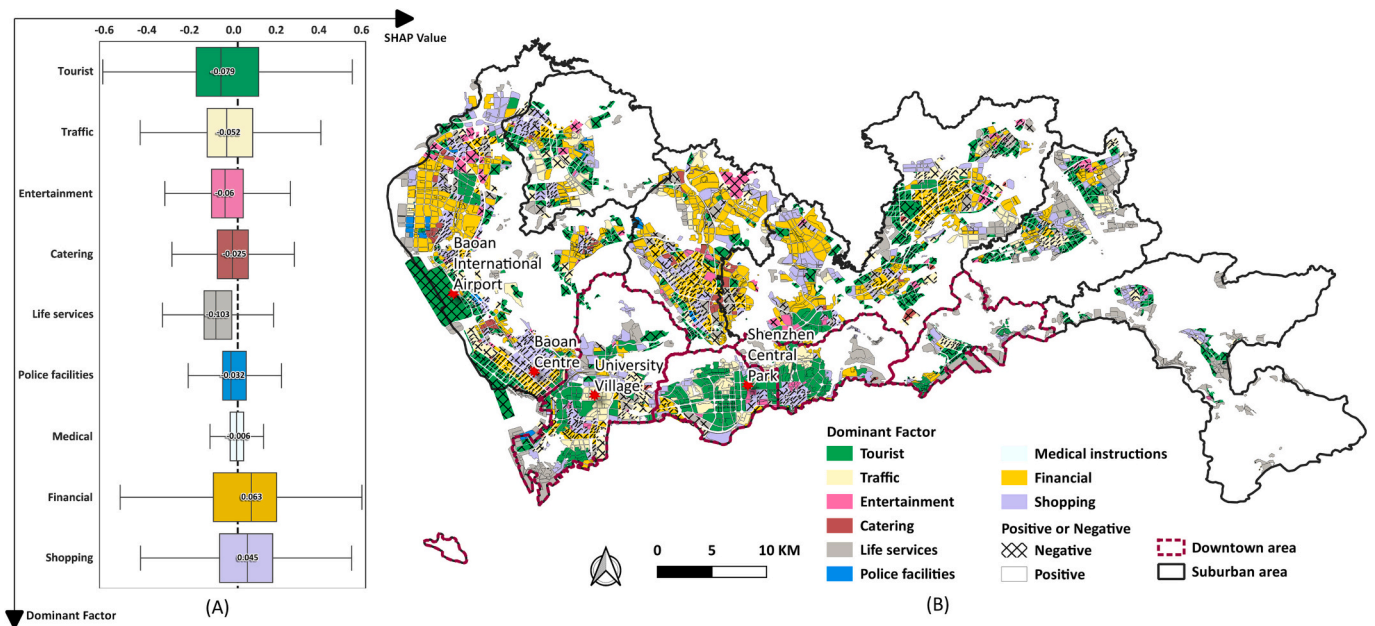


Fig. 9. Distribution of SHAP values and main drivers for all land parcels in Shenzhen: (A) shows the statistical analysis of SHAP values for all land parcels with the same characteristics; the SHAP Value axis indicates the magnitude of SHAP values, and the Dominant Factor axis indicates each type of POI characteristic that affects the risk of pickpocketing; (B) shows the drivers that have the greatest impact on the risk of pickpocketing in the parcel, which is obtained based on the average absolute magnitude of SHAP.

proportion of migrant residents.

Conversely, 26.8 % of the area exhibited low-low aggregation regions, mainly distributed outside the central city. Notably, these regions typically had better built environments and overall policing, evident in areas such as the central southern region of Baoan district, the science and technology industrial area in the south-central Longhua district, and the northeastern industrial area in Longgang district (Fig. 8(D)-(F)). Our study findings indicate socioeconomic conditions and the built environment in neighbouring regions greatly influence the spatial pattern of crime risk. The results are consistent with the hypothesis proposed by Sparks (2011a).

4.4. Explainable spatial distribution of pickpocketing risks

After fitting the relationship between each type of POI feature and the risk of pickpocketing crime, we calculated the SHAP values for each type of POI feature (Fig. 9(A)). Residents' routine activity was mainly carried out in five facilities, namely Traffic, Entertainment, Catering, Shopping, Financial, and Life services facilities (Boivin, 2018). The results indicate that routine activity was the most important factor influencing pickpocketing crime risk and positively correlates with the risk of pickpocketing crime. Compared to Tourist (-0.052), Medical Institutions (-0.003), and Police (-0.009), routine activity had the greatest impact on the risk of pickpocketing crime with an overall SHAP value of 0.079. According to the resident's routine activity theory (Cohen & Felson, 1979), Routine activity facilities provide criminals and potential targets that are prone to criminal activity. Tourist characteristics had a negative impact on pickpocketing crime risk scores, with a negative median SHAP value (-0.079). The Tourist feature reduced the risk score in many samples, which indicates that tourism had a high potential to reduce the risk of pickpocketing crime in urban areas. The results are consistent with the findings of (Bogar & Beyer, 2016) that urban landscape features are associated with reduced crime rates.

The present study investigated the determinants of pickpocketing crime risk on all parcels in Shenzhen Shenzhen, revealing insights into the spatial heterogeneity of crime risks (Fig. 9(B)). The influence of routine activities on the risk of pickpocketing crime was found to be spatially heterogeneous. In the downtown area, routine activity facilities were observed to positively affect the risk of pickpocketing (average SHAP value: 0.184). For example, in the vicinity of Shenzhen University Town, where traffic was the main driver, urban residents moved around for work and education purposes, and the population was more mobile, increasing the risk of pickpocketing crime in the area. However, residents' routine activity facilities in suburban areas negatively impact crime risk (average SHAP value: -0.118). For example, Shopping and Financial features dominated the dominant factors near the central area of Baoan district, which reduced the crime risk of the area. This may associate with increased guardianship (Boivin, 2018), which stabilizes social order. These results support the hypothesis that routine activity may increase or decrease criminal activity (Boivin, 2018).

The present study revealed the heterogeneous effect of tourist attractions on the risk of pickpocketing crime. Tourist (average SHAP value: 0.098) had a predominantly positive effect on crime risk in economically developed urban areas. For instance, Shenzhen Central Park witnessed an increased risk of pickpocketing due to the congregation of many tourists, making tourism the main driver of crime risk in the area (Zhong et al., 2011). In contrast, tourists in suburban areas (average SHAP value: -0.119) mainly negatively affect crime risk. For example, crime risk in Baoan International Airport was driven by tourism, but it reduced the risk of pickpocketing crime in the area. This could be attributed to the presence of a large area of public green spaces near the airport, which stabilizes social order and reduces the risk of pickpocketing crime (Jennings & Bamkole, 2019). Our study shows the uncertainty of Tourist's effect on crime risk in the region due to socioeconomic influences, which is in line with the findings of (Groff & McCord, 2012).

5. Discussion

The extraction of crime information from street view images reflecting the built environment is essential for urban governance and crime risk analysis. However, the number of street view images labeled as crime occurred is often very less. This issue is particularly true in China since the most reliable crime data source is the Chinese judgment documents, which do not contain all criminal cases. This study is an active attempt to extract crime risk through urban built environments using spatially sparse crime data. To achieve this, we adopted an alternative approach by evaluating the distribution of pickpocketing crimes based on OCC-based anomaly detection and street view images. In addition, it is the first exploration of the relationship between the built environment and pickpocketing crime risk in a large Chinese city. Previous studies have analyzed cities' physical environment and socioeconomic characteristics as reflected in street view images in several U. S. cities and explored their relationship with urban crime rates (He et al., 2017; Zhang et al., 2021).

5.1. Interpretation of the findings

In this study, we developed a Crime Anomaly Detection based on the Street View (CADSV) model, which can effectively extract deep semantic information from massive street view for assessing pickpocketing risks. Our results show that the street view images can accurately reflect the city's physical environment and provide reliable assessments of the risk of pickpocketing crimes, as confirmed by the high accuracy of the CADSV model (AUC = 0.921, recall = 0.816). Through comparative and explainable analysis, we obtained a micro-scale pickpocketing risk distribution map of the study area and confirmed the reliability of the results.

Our findings reveal that pickpocketing crime in Chinese megacities exhibits strong spatial autocorrelation, consistent with previous studies' observations that crime tends to be concentrated in small areas (Groff et al., 2010; Weisburd et al., 2004). The observed decrease in crime risk decreases with distance from the downtown provides quantitative support for the social disorganization theory proposed by Shaw et al. (1942). In the downtown area, pickpocketing is a high prevalence and aggregation of pickpocketing crime are significant due to dense human traffic, making it challenging to manage. Simultaneously, in the course of ongoing urban expansion due to population and economic growth, the influx of migrant workers and the gradual deterioration of the physical environment in older urban areas (Li et al., 2014) have become the dominant drivers of the high prevalence of pickpocketing crimes in urban village areas (Liu, 2010). In contrast, the suburbs, and industrial parks outside the central city, where a large number of immigrants live and work (Roitman & Phelps, 2011), display a better built environment and exhibit low-low aggregation of pickpocketing crimes. These findings highlight the complexity of the impact of urban function on pickpocketing crime. They also confirm that confirm that the complex roles of the urban physical environment, neighborhood socioeconomics, and migrant population all play a significant role in shaping the spatial distribution of pickpocketing crime risk in Chinese cities (Sparks, 2011b).

This study also utilized the Random Forest and SHAP model to interpretively analyze the relationship between pickpocketing crime, urban function, and urban environment at the microscopic scale. The proposed model achieved high accuracy ($R^2 = 0.455$) and reliability (RMSE = 0.016) by employing urban functions to fit pickpocketing risk, thus quantitatively confirming the crucial role of different urban functions in shaping regional pickpocketing risk. Consistent with the routine activity theory proposed by Cohen and Felson (1979) and the crime pattern theory proposed by Brantingham and Brantingham (2013), our findings highlight that routine activity in the central city is a critical factor that enhances pickpocketing risks. Furthermore, we observed that the high intensity of economic activity in commercial areas contributes

significantly to crime incidence in China.

The results demonstrate that both the built environment information and urban functional information captured in street view images play a role in shaping crime incidence. Regarding the built environment, areas with high-density urban buildings and visually chaotic and disorganized surroundings are more susceptible to pickpocketing crimes, while areas with better-organized environments have lower risks. Concerning urban functional zones, densely populated areas that are challenging to fully secure and where high-value items are prevalent pose a higher risk of pickpocketing crimes. Examples include high-traffic attractions, shopping centres, isolated factories or residential areas, schools with a high number of minors, among others. We can use these findings to guide urban planning and security management. For example, in high-traffic areas such as commercial centres and tourist attractions, surveillance and security forces can be strengthened in advance to reduce the threat of crime by installing additional warning signs and alarm facilities based on unusual risk situations reflected in the street view images. In important areas such as residential and school zones, residents and students can be encouraged to exercise extra vigilance towards high-value property. Additionally, open spaces like parks and attractions could benefit from increased police patrols and surveillance equipment deployments in crowded areas can be increased to improve security perceptions and prevent pickpocketing and other crimes from happening. Furthermore, functional areas may also reflect differences in social structures. For instance, the transport facilities present a higher risk of crime in the city centre than in other areas, which may have complex links to social phenomena such as frequent movements of movement of people, conflicts arising from intricate social structures, and the allocation of police forces. In the long term, governments should to promote social development and long-term security by improving the social structure and raising the level of the economy.

The above discussion underscores the multifaceted nature of pickpocketing crime in China's megacity, which cannot be explained by a single theory of crime. The occurrence of such crime is influenced by a combination of regional economic development, urban physical environment, and routine activity, among other factors. Moreover, interpretable analyses have uncovered complex spatial heterogeneity in the drivers of pickpocketing crime across China's megacity. Routine activity exerts a positive impact on areas with intense human activity areas in downtown regions but a negative impact in suburban areas. Similarly, tourist activity positively affects crime risk in urban centers but has a negative effect in the suburbs. These findings indicate that the dichotomy of China's urban-rural structure, characterized by the differences in physical space, industrial infrastructure, and economic composition across regions (Ann et al., 2015; Long et al., 2016), gives rise to significant variations in the underlying drivers of crime.

As a developing country, China faces the challenge of operating with a relatively constrained police force and fewer police services available per capita compared to developed countries, resulting in inadequate law enforcement resources to combat pickpocketing crimes (Hyland & Davis, 2019; Wang et al., 2014). To enhance policing effectiveness, our findings suggest deploying police patrols strategically high-risk areas for pickpocketing crimes while implementing video surveillance systems in urban villages, shopping centres, and economic activity centers. Furthermore, increasing anti-pickpocketing campaigns at daily activity locations such as bus stops and metro stations could raise residents' security awareness and contribute to reducing pickpocketing risks. Our study further highlights that urban villages with significant migrant populations are at greater risk of pickpocketing crime. Therefore, improving service facilities in urban villages and providing more employment opportunities may aid in enhancing urban policing efforts.

There are difficulties in mapping the real spatial pattern of crime risk due to the specificity of data collection for the judgment document. The anomaly detection model in this study learns from sparse data about hidden crime risks and finds a spatial mismatch between the number of crime events and the risk of the area. Current policy-making authorities

quantify the level of policing in an area mostly based on government survey data, such as judgment documents. Conclusions based on such data may therefore lead to problems such as misallocation of public resources and misguided business investments. Our findings may provide support to government policy makers or commercial investors.

5.2. The spatial mismatch between the judgment document and mapped crime risk

Our study has revealed a spatial mismatch between crime risk and the original crime data obtained from judgment document, as depicted in Fig. 10(A). This disparity is a tangible manifestation of the sparse and biased sampling problem that this research has explored. Specifically, the spatial distribution of data collected through sentencing instruments is exceedingly sparse and closely associated with factors such as population density and law enforcement efficiency, rendering it difficult to accurately reflect the actual spatial pattern of crime risk.

In practice, acquiring a judgment document involves a lengthy process comprising three stages: (1) commission of a crime and successful theft; (2) notification of the police by the victim or public body, leading to the opening of a case and arrest of the suspect; and (3) filing of a case and commencement of prosecution against the accused by the victim or public body in a court of law. Consequently, the data we collect for each judgment paper represents not only the occurrence of a crime, but also the diligence of the court and the efficiency of the police in executing the case. Regarding the distribution of crimes, since not all criminal incidents go through the aforementioned process, the sentencing paper data can only serve as a sparse sample point and cannot directly depict the full scope of criminal activities.

Regarding the sparsity of sampled data, Table 1 presents the area covered by crime points and the number of street view images under varying buffer distances. The results reveal that with a buffer distance of 200 m, only 3 % of the region is labeled as crime-related, with an average of 4.17 street view images per buffer. If a shorter buffer distance of 50 m is chosen, then merely 0.2 % of the area is covered, and each buffer can only include 0.31 street image. Given the limited proportion of relevant data, it becomes arduous to offer a comprehensive and accurate spatial pattern at a global level.

With regards to the biased nature of the data, we counted the judgment instruments for all cases (including cases in which the location is not publicly available) in each administrative region of Shenzhen in 2018 based on the data provided by the Judgment Instruments website (<https://wenshu.court.gov.cn/>), as shown in Fig. 10(B). The figure depicts darker colors indicating a higher total number of judgment instruments within each respective region. It is evident that the aggregation level of crime points obtained through our opportunity judgment instruments closely aligns with the number of judgment instruments generated by courts in each region.

In order to further validate our previous assertion concerning human activity, we have gathered Real-time Tencent user density (RTUD) data. Tencent is one of the largest internet companies in China, with a user base exceeding 800 million individuals utilizing its diverse range of internet services. Through Tencent Maps or WeChat, when users engage in location-related activities, their relevant location information is recorded, enabling RTUD data to capture population distribution during specific periods. The raw data is stored as a raster image format comprising of 24 bands that represent each hour of the day. This study utilizes an overlay of the 24-h average change in population density over weekdays to generate a graph (He et al., 2020). A correlation can be observed between higher crime spots and elevated levels of people's activities, once again affirming the notion that data acquisition does not accurately reflect the complete volume of criminal incidents.

This study has examined areas of mismatch to establish a connection between Fig. 10 (A) and Fig. 8. Specifically, in Figs. 10 (A-a) and Fig. 8 (C), it is apparent that despite the limited sample of pickpocketing crime events in Tong Mei Urban Village, the village exhibits a high risk of

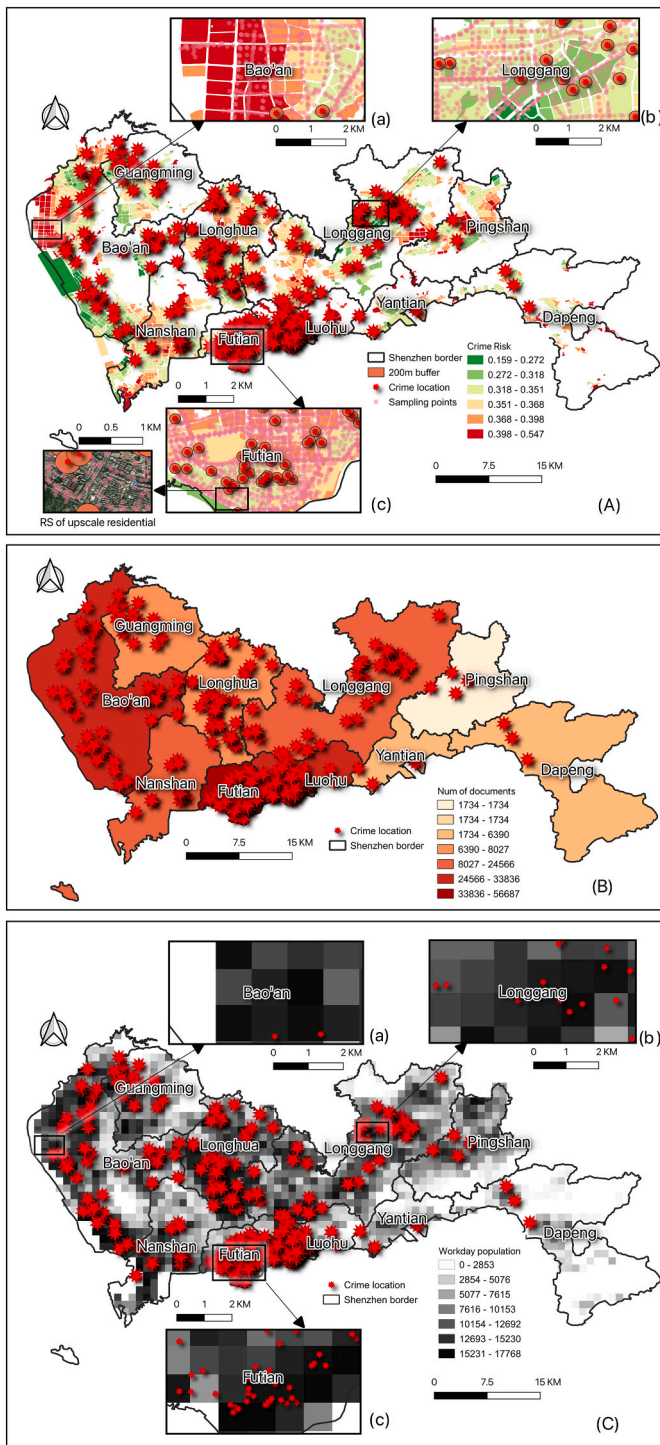


Fig. 10. The overlap of pickpocketing events between the judgment document and: (A) the crime risk distribution; (B) the total number of judgment documents of each district; (C) The average population distribution of Shenzhen during the working day.

criminal activity due to its inadequate infrastructure and chaotic building structure. Conversely, Fig. 10 (A-b) and Fig. 8 (E) demonstrate that emerging technology industrial parks and factories have a lower risk of crime, attributed to their well-organized and tidy environment. Further, Fig. 10 (A-c) corresponds to Fig. 8 (B), where the Futian Central Business District displays a medium-high risk of criminal activity due to the concentration of wealth. However, there are also areas with a low risk of criminal activity, as seen in Fig. 10(A-c). These regions are

Table 1

The area and images that are covered by crime event according to different buffer.

Buffer range	50 m	100 m	150 m	200 m	250 m	300 m
Area percentage covered by buffer zones	0.2 %	0.8 %	1.9 %	3.0 %	5.4 %	7.8 %
Average number of images covered by one buffer	0.31	1.36	2.51	4.17	6.69	9.45

primarily located in residential areas within the city center, which may be due to enhanced security facilities and the higher quality of residents.

5.3. Limitation and future works

The present study does have some limitations that must be acknowledged. First, the objective of this study is to utilize sparse crime data to reflect implicit crime risks in urban built environments using the technique of anomaly detection. However, it is important to note that this study conducted a crime risk analysis rather than an estimation of actual crime rates. There are interactions and complex causal relationships between risk and crime rates. Understanding these relationships is critical to developing effective crime prevention and governance strategies, and requires the integration of multiple social, economic, cultural and individual dimensions. Future studies may collaborate with law enforcement agencies to analyze the relationship between street view images and real-time alarm data utilizing the framework proposed in this paper.

The second limitation of this study pertains to the crime data used in model design. The premise assumption of this study is that crime information can be viewed as an outlier in the urban landscape. Therefore, we first removed street view images spatially associated with a crime based on judgment documents. After that, we randomly selected a certain number of street view images for training purposes so as to obtain a normal feature vector. However, we cannot guarantee that no crime has occurred in those areas since judgment instruments may not always contain all relevant data. Our hypothesis was that by using deep learning for feature extraction using numerous images, we could eliminate the influence of crime information could be eliminated as much as possible. The precision validation results also show such effectiveness. To further improve the accuracy, subsequent studies should take into account the more prior knowledge and eliminate as much as possible the street view images where crime may be present to obtain the most effective feature vector. Moreover, exploring the impact of different street view image acquisition intervals on the results would be valuable. More frequent street view sampling has the potential to yield better results, and in the field of crime, how to choose the most suitable analysis interval is also a topic worth exploring (Ramos et al., 2021). This study has successfully constructed a framework illustrating the feasibility of anomaly detection for exploring crime risk. Subsequent research endeavors can build upon this framework to delve deeper into this topic.

The third limitation is we utilized POI data to interpret the result of crime risk mapping. Prior research has found that POI data can effectively reflect the characteristics of socioeconomic structure (Yao et al., 2017). Furthermore, we have carried out some work to demonstrate the strong relationship between street view images and several social and environmental factors such as urban economic level and urban population structure (Wang et al., 2021; Yao et al., 2021b). However, POI data and street view images can only be proxy variables of socioeconomic characteristics and urban environment. To explain the risk mapping result more accurately, future studies will introduce more detailed census, travel survey, and trajectory data. Additionally, econometric models may be used to analyze the temporal and spatial

correlations between the multiple urban structures and criminal behaviours at the micro-scale. Finally, follow-up research can also expand the research scale by obtaining global street view datasets and analyzing the similarities and differences of criminal behavior drivers in different cities worldwide.

The solution proposed in this study carries substantial practical value, as street view images are easy to obtain, models can be easily migrated to other areas, and it can be utilized by lay users to quickly comprehend crime risk levels in a particular area of the city. For instance, we can develop a mobile application that enables users to swiftly assess safety status of a city neighborhood. Visitors arriving in an unfamiliar city can rapidly determine whether a specific alley is safe. While police crime statistics are typically the most trustworthy in such cases, official data may not always cover the entire area. In such scenarios, our approach can assist users in identifying and mitigating potential safety concerns.

6. Conclusion

This study aims to propose a solution for extracting crime risk information from the built environment using the limited crime-labeled street view images. We also try to prove the association between the human perception of the built environment and urban pickpocketing crimes in China. To achieve these objectives, we propose a pickpocketing risk assessment model that combines deep anomaly detection techniques to reveal the crime risk from street view images. The SHAP was introduced to conduct an interpretable analysis of urban functions and crime risks. Through spatial distribution analysis of pickpocketing crimes based on judicial documents, our proposed CADSV model accurately and reliably maps out a micro-scale pickpocketing risk distribution in Shenzhen. Our results indicate street view images can effectively assess pickpocketing crime risk in Chinese cities, and the crime risk has a strong spatial autocorrelation. Moreover, we demonstrate that pickpocketing crime in China is driven by complex factors such as regional economic development, physical urban environment, and daily activities. These findings provide valuable insights for policing deployment and city management strategies. Nonetheless, this study does not discuss the association between crime risk and actual crime rates. Moreover, the inclusion of finer-scale geographic big data could be considered to identify crime-related street view images in anomaly detection, thereby offering more prior knowledge about crime risk. Furthermore, a more comprehensive interpretation of the crime risk mapping results is necessary to analyze the correlation between various urban structures and criminal behavior at finer spatial and temporal scales.

CRedit authorship contribution statement

Yao Yao: Conceptualization, Methodology, Writing- Reviewing and Editing. **Anning Dong:** Data curation, Writing- Original draft preparation. **Zhiqian Liu:** Visualization, Investigation. **Ying Jiang:** Supervision. **Zijin Guo:** Software, Validation. **Junyi Cheng:** Writing- Reviewing and Editing. **Qingfeng Guan:** Writing- Reviewing and Editing. **Peng Luo:** Conceptualization, Methodology, Writing- Original draft preparation.

Fundings

This work was supported by the National Key Research and Development Program of China (Grant No. 2019YFB2102903), the National Natural Science Foundation of China (Grand No. 41801306 and 42171466), Alibaba Innovative Research Project (No. 20228670), and China Scholarship Council.

Declaration of competing interest

The authors declare that they have no known competing financial

interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

- Ann, T. W., et al. (2015). The key causes of urban-rural conflict in China. *Habitat International*, 49, 65–73.
- Ashihara, Y., & Riggs, L. E. (1983). *The aesthetic townscape*. Cambridge: MIT Press.
- Bernasco, W., Block, R., & Ruiter, S. (2013). Go where the money is: Modeling street Robbers' location choices. *Journal of Economic Geography*, 13(1), 119–143.
- Bernasco, W., Ruiter, S., & Block, R. (2017). Do street robbery location choices vary over time of day or day of week? A test in Chicago. *Journal of Research in Crime and Delinquency*, 54(2), 244–275.
- Bogar, S., & Beyer, K. M. (2016). Green space, violence, and crime: A systematic review. *Trauma, Violence & Abuse*, 17(2), 160–171.
- Boivin, R. (2018). Routine activity, population (S) and crime: Spatial heterogeneity and conflicting propositions about the neighborhood crime-population link. *Applied Geography*, 95, 79–87.
- Bouma, H., et al., 2014. Automatic detection of suspicious behavior of pickpockets with track-based features in a shopping mall. In Optics and Photonics for Counterterrorism, Crime Fighting, and Defence X; and Optical Materials and Biomaterials in Security and Defence Systems Technology XI (International Society for Optics and Photonics), 92530F.
- Brantingham, P., & Brantingham, P. (2013). Crime pattern theory. In *Environmental Criminology and Crime Analysis (Willan)* (pp. 100–116).
- Brunton-Smith, I., et al. (2023). Estimating the reliability of crime data in geographic areas (CrimRxiv).
- Buil-Gil, D., Moretti, A., & Langton, S. H. (2022). The accuracy of crime statistics: Assessing the impact of police data Bias on geographic crime analysis. *Journal of Experimental Criminology*, 18(3), 515–541.
- Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, 36(3), 4626–4636.
- Byun, G., & Kim, Y. (2022). A street-view-based method to detect urban growth and decline: A case study of midtown in Detroit, Michigan, USA. *PLoS One*, 17(2), Article e263775.
- Cai, T., & Xin, Y. (2019). Child trafficking in China: Evidence from sentencing documents. *International Journal of Population Studies*, 4(2), 1–11.
- Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), 15.
- Cohen, L. E., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. *American Sociological Review*, 588–608.
- Cohen, N., & Hoshen, Y. (2020). Sub-image anomaly detection with deep pyramid correspondences. *arXiv e-prints*. arXiv:2005.02357.
- Deshoteles, T. (2013). The declining criminal arts-the pickpocket. *The International Journal of Crime, Criminal Justice and Law/Serials Publication*, 8, 69–76.
- Ding, N., & Zhai, Y. (2021). Crime prevention of bus pickpocketing in Beijing, China: Does air quality affect crime? *Security Journal*, 34(2), 262–277.
- Du, F., et al. (2020). Predictive mapping with small field sample data using semi-supervised machine learning. *Transactions in GIS*, 24(2), 315–331.
- Fan, N., et al. (2020). Digital soil mapping over large areas with invalid environmental covariate data. *ISPRS International Journal of Geo-Information*, 9(2), 102.
- Fan, Z., et al. (2021). Evaluation of urban crime risk based on agent simulation model. In *2021 4th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE) IEEE* (pp. 648–651).
- Farrell, G., & Bouloukos, A. C. (2001). International overview: a cross-national comparison of rates of repeat victimization. *Crime Prevention Studies*, 12, 5–26.
- Fernández-Delgado, M., et al. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15(1), 3133–3181.
- Giménez-Santana, A., Caplan, J. M., & Drawwe, G. (2018). Risk terrain modeling and socio-economic stratification: Identifying risky places for violent crime victimization in Bogotá, Colombia. *European Journal on Criminal Policy and Research*, 24(4), 417.
- Groff, E., & McCord, E. S. (2012). The role of neighborhood parks as crime generators. *Security Journal*, 25(1), 1–24.
- Groff, E. R., Weisburd, D., & Yang, S. (2010). Is it important to examine crime trends at a local "Micro" level?: A longitudinal analysis of street to street variability in crime trajectories. *Journal of Quantitative Criminology*, 26(1), 7–32.
- Hajela, G., Chawla, M., & Rasool, A. (2021). Crime hotspot prediction based on dynamic spatial analysis. *ETRI Journal*, 43(6), 1058–1080.
- He, J., et al. (2020). Accurate estimation of the proportion of mixed land use at the street-block level by integrating high spatial resolution images and geospatial big data. *IEEE Transactions on Geoscience and Remote Sensing*, 59(8), 6357–6370.
- He, L., Páez, A., & Liu, D. (2017). Built environment and violent crime: An environmental audit approach using Google street view. *Computers, Environment and Urban Systems*, 66, 83–95.
- Helbich, M., et al. (2019). Using deep learning to examine street view green and blue spaces and their associations with geriatric depression in Beijing, China. *Environment International*, 126, 107–117.

- Hengl, T., et al. (2018). Random Forest as a generic framework for predictive modeling of spatial and Spatio-temporal variables. *PeerJ*, 6, Article e5518.
- Hipp, J. R., et al. (2019). Using social media to measure temporal ambient population: Does it help explain local crime rates? *Justice Quarterly*, 36(4), 718–748.
- Hipp, J. R., et al. (2021). Measuring the built environment with Google street view and machine learning: Consequences for crime on street segments. *Journal of Quantitative Criminology*, 1–29.
- Hossain, S., et al. (2020). Crime prediction using spatio-temporal data. In *International Conference on Computing Science, Communication and Security (Springer)* (pp. 277–289).
- Hu, Y., & Han, Y. (2019). Identification of urban functional areas based on POI data: A case study of the Guangzhou economic and technological development zone. *Sustainability*, 11(5), 1385.
- Hu, Y., et al. (2018). A Spatio-temporal kernel density estimation framework for predictive crime hotspot mapping and evaluation. *Applied Geography*, 99, 89–97.
- Hyland, S. S., & Davis, E. (2019). Local police departments, 2016: Personnel. In *Washington, DC: Bureau of Justice Statistics (BJS) US Dept of justice, Office of Justice Programs*. Bureau of Justice Statistics.
- Jennings, V., & Bamkole, O. (2019). The relationship between social cohesion and urban green space: An avenue for health promotion. *International Journal of Environmental Research and Public Health*, 16(3), 452.
- Jing, F., et al. (2021). Assessing the impact of street-view greenery on fear of neighborhood crime in Guangzhou, China. *International Journal of Environmental Research and Public Health*, 18(1), 311.
- Kadar, C., Iria, J. E., & Cvijikj, I. P. (2016). Exploring foursquare-derived features for crime prediction in new York City. *KDD-Urban Computing WS*, 16, 10–1145.
- Kang, Y., et al. (2020). A review of urban physical environment sensing using street view imagery in public health studies. *Annals of GIS*, 26(3), 261–275.
- Khosla, P., et al. (2020). Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33, 18661–18673.
- Kim, Y., & Kim, H. K. (2021). Cluster-based deep one-class classification model for anomaly detection. *Journal of Internet Technology*, 22(4), 903–911.
- Kim, Y., & Kim, Y. (2022). Explainable heat-related mortality with random Forest and SHapley additive exPlanations (SHAP) models. *Sustainable Cities and Society*, 79, 103677.
- Lafree, G., & Birkbeck, C. (2010). The neglected situation: A cross-national study of the situational characteristics of crime. *Criminology*, 29(1), 73–98.
- Li, L. H., et al. (2014). Redevelopment of Urban Village in China—a step towards an effective urban policy? A case study of Liede Village in Guangzhou. *Habitat International*, 43, 299–308.
- Li, Y., et al. (2016). Supplemental sampling for digital soil mapping based on prediction uncertainty from both the feature domain and the spatial domain. *Geoderma*, 284, 73–84.
- Ling, C. X., Huang, J., & Zhang, H. (2003). AUC: A statistically consistent and more discriminating measure than accuracy. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (Acapulco, Mexico: Morgan Kaufmann Publishers Inc.)* (pp. 519–524).
- Liu, Z. G. (2010). The mechanism of crimes in villages-in-city: A case study of “T Village” in Shenzhen. *Criminal Research*, 17(6), 64–71.
- Long, H., et al. (2016). The allocation and Management of Critical Resources in rural China under restructuring: Problems and prospects. *Journal of Rural Studies*, 47, 392–412.
- Lundberg, S., & Lee, S. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (Curran Associates Inc.)* (pp. 4768–4777).
- Massoli, F. V., Falchi, F., Kantarci, A., Akti, Ş., Ekenel, H. K., & Amato, G. (2021). MOCCA: Multilayer one-class classification for anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*, 33(6), 2313–2323.
- Meyer, D. R. (2016). Shenzhen in China's financial center networks. *Growth and Change*, 47(4), 572–595.
- Miao, L., et al. (2016). Add to favorite get latest update discussion on the improvement of the crime of medical accident in China: Based on the analysis of the cases of China judgments online and Jianxue Li case. *Medicine and Philosophy: A*, 37(12), 3.
- Minhas, M. S., & Zelek, J. (2019). Anomaly detection in images. *arXiv e-prints*, 1905–13147.
- Nussbaum, M., et al. (2018). Evaluation of digital soil mapping approaches with large sets of environmental covariates. *Soil*, 4(1), 1–22.
- Oswald, M., et al. (2018). Algorithmic risk assessment policing models: Lessons from the Durham HART model and ‘experimental’ proportionality. *Information & Communications Technology Law*, 27(2), 223–250.
- Perera, P., Oza, P., & Patel, V. M. (2021). One-class classification: A survey. *arXiv e-prints*. arXiv:2101.03064.
- Ramos, R. G., et al. (2021). Too fine to be good? Issues of granularity, uniformity and error in spatial crime analysis. *Journal of Quantitative Criminology*, 37, 419–443.
- Reiss, T., et al. (2021). PANDA: Adapting pretrained features for anomaly detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2806–2814).
- Reiss, T., & Hoshen, Y. (2023, June). Mean-shifted contrastive loss for anomaly detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2), 2155–2162.
- Rodriguez-Galiano, V. F., et al. (2012). An assessment of the effectiveness of a random Forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 67, 93–104.
- Roitman, S., & Phelps, N. (2011). Do gates negate the City? Gated Communities’ contribution to the urbanisation of suburbia in Pilar, Argentina. *Urban Studies*, 48(16), 3487–3509.
- Ruff, L., et al. (2018). Deep one-class classification. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research: PMLR)* (pp. 4393–4402).
- Rumi, S. K., Luong, P., & Salim, F. D. (2019). Crime rate prediction with region risk and movement patterns. *CoRR*. abs/1908.02570.
- Sabokrou, M., Khalooei, M., Fathy, M., & Adeli, E. (2018). Adversarially learned one-class classifier for novelty detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3379–3388).
- Schisterman, E. F., et al. (2005). Optimal cut-point and its corresponding Youden index to discriminate individuals using pooled blood samples. *Epidemiology*, 16(1), 73–81.
- Schlegl, T., et al. (2017). Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International Conference on Information Processing in Medical Imaging (Cham: Springer International Publishing)* (pp. 146–157).
- Shaw, C. R., et al. (1942). Juvenile delinquency and urban areas: A study of rates of delinquents in relation to differential characteristics of local communities in American cities. *University of Chicago Press*, 49(1), 100–101.
- Sparks, C. S. (2011a). Violent crime in San Antonio, Texas: An application of spatial epidemiological methods. *Spatial and Spatio-temporal Epidemiology*, 2(4), 301–309.
- Sparks, C. S. (2011b). Violent crime in San Antonio, Texas: An application of spatial epidemiological methods. *Spatial and spatio-temporal epidemiology*, 2(4), 301–309.
- Steffensmeier, D., Zhong, H., & Lu, Y. (2017). Age and its relation to crime in Taiwan and the United States: Invariant, or does cultural context matter? *Criminology*, 55(2), 377–404.
- Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3), 647–665.
- ToppiReddy, H. K. R., Saini, B., & Mahajan, G. (2018). Crime Prediction & Monitoring Framework Based on Spatial Analysis. *Procedia Computer Science*, 132, 696–705.
- Wang, F., & Liu, H. (2021). Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2495–2504).
- Wang, R., et al. (2019a). Using street view data and machine learning to assess how perception of neighborhood safety influences urban Residents’ mental health. *Health & Place*, 59, 102186.
- Wang, R., et al. (2019b). The linkage between the perception of Neighbourhood and physical activity in Guangzhou, China: Using street view imagery with deep learning techniques. *International Journal of Health Geographics*, 18(1), 18.
- Wang, R., et al. (2021). The distribution of greenspace quantity and quality and their association with Neighbourhood socioeconomic conditions in Guangzhou, China: A new approach using deep learning method and street view images. *Sustainable Cities and Society*, 66, 102664.
- Wang, Y., et al. (2014). Stress, burnout, and job satisfaction: Case of police force in China. *Public Personnel Management*, 43(3), 325–339.
- Webb, V. J., et al. (2011). A comparative study of youth gangs in China and the United States: Definition, offending, and victimization. *International Criminal Justice Review*, 21(3), 225–242.
- Weisburd, D. (2015). The law of crime concentration and the criminology of place. *Criminology*, 53(2), 133–157.
- Weisburd, D., et al. (2004). Trajectories of crime at places: A longitudinal study of street segments in the City of Seattle. *Criminology*, 42(2), 283–322.
- Wilson, J., & Kelling, G. L. (1982). Broken windows: The police and neighbourhood safety. *The Atlantic Monthly*, 249, 29–38.
- Xiao, J., & Zhou, X. (2020). Crime exposure along my way home: Estimating crime risk along personal trajectory by visual analytics. *Geographical Analysis*, 52(1), 49–68.
- Xie, H., Liu, L., & Yue, H. (2022). Modeling the effect of streetscape environment on crime using street view images and interpretable machine-learning technique. *International Journal of Environmental Research and Public Health*, 19(21), 13833.
- Yao, Y., et al. (2017). 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.
- Yao, Y., et al. (2019). A human-machine adversarial scoring framework for urban perception assessment using street-view images. *International Journal of Geographical Information Science*, 33(12), 2363–2384.
- Yao, Y., et al. (2021a). Delineating urban job-housing patterns at a parcel scale with street view imagery. *International Journal of Geographical Information Science*, 35(10), 1927–1950.
- Yao, Y., et al. (2021b). Delineating urban job-housing patterns at a parcel scale with street view imagery. *International Journal of Geographical Information Science*, 1–24.
- Yao, Y., et al. (2022). Assessing myocardial infarction severity from the urban environment perspective in Wuhan, China. *Journal of Environmental Management*, 317, 115438.
- Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1), 32–35.
- Yue, H., et al. (2022). Detecting people on the street and the streetscape physical environment from Baidu street view images and their effects on community-level street crime in a Chinese City. *ISPRS International Journal of Geo-Information*, 11, 151.
- Zhang, F., et al. (2018). Measuring human perceptions of a large-scale urban region using machine learning. *Landscape and Urban Planning*, 180, 148–160.
- Zhang, F., et al. (2020). Uncovering inconspicuous places using social media check-ins and street view images. *Computers, Environment and Urban Systems*, 81, 101478.
- Zhang, F., et al. (2021). “perception Bias”: Deciphering a mismatch between urban crime and perception of safety. *Landscape and Urban Planning*, 207, 104003.
- Zhang, G. (2022). Mitigating spatial bias in volunteered geographic information for spatial modeling and prediction. In *New Thinking in GIScience* (pp. 179–190). Springer.
- Zhang, G., & Zhu, A. (2018). The representativeness and spatial bias of volunteered geographic information: A review. *Annals of GIS*, 24(3), 151–162.

- Zhang, G., & Zhu, A. (2019a). A representativeness heuristic for mitigating spatial bias in existing soil samples for digital soil mapping. *Geoderma*, 351, 130–143.
- Zhang, G., & Zhu, A. (2019b). A representativeness-directed approach to mitigate spatial bias in VGI for the predictive mapping of geographic phenomena. *International Journal of Geographical Information Science*, 33(9), 1873–1893.
- Zhang, X., et al. (2022). Interpretable machine learning models for crime prediction. *Computers, Environment and Urban Systems*, 94, 101789.
- Zhao, X., & Tang, J. (2017). Modeling temporal-spatial correlations for crime prediction. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 497–506). Singapore: Association for Computing Machinery.
- Zhong, H., et al. (2011). Spatial analysis for crime pattern of metropolis in transition using police records and GIS: A case study of Shanghai, China. *International Journal of Digital Contents Technology and Its Applications*, 5(2), 93–105.
- Zhu, A. X., et al. (2018). Spatial prediction based on third law of geography. *Annals of GIS*, 24(4), 225–240.