UNCERTAINTY ANALYSIS AND DATA QUALITY ASSURANCE IN SPATIAL BIG DATA

PENGFEI CHEN

PhD

The Hong Kong Polytechnic University

This programme is jointly offered by The Hong Kong Polytechnic University and Wuhan University

2020

The Hong Kong Polytechnic University

Department of Land Surveying and Geo-Informatics

Wuhan University

School of Remote Sensing and Information Engineering

Uncertainty analysis and data quality assurance in spatial big data

Pengfei CHEN

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

December 2019

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____ (Signed)

CHEN Pengfei (Name of student)

Uncertainty analysis and data quality assurance in spatial big data

Abstract

Since the term ‘big data’ was coined for the first time in 2005, it has unleashed a worldwide evolution in scientific research and business. Spatial big data (SBD), the big data associated with geographical information, are one of the most valuable products in modern science motivated by the rapid development of smart technology and sensor technology. Generally, SBD can be classified as earth observation data and human activity data. Thus far, SBD has stimulated a continuous wave of innovations in a wide range of disciplines, such as geoscience, urbanology and environmental science.

Uncertainty has been long recognised as an essential element affecting the entire process of spatial data production and analysis. Inappropriate uncertainty management can result in misleading knowledge and cause tremendous losses. Over the past decades, considerable efforts have been made to develop theories and methods for uncertainties in spatial data and analytics. However, given the continuously increasing complexity and volume of SBD, traditional uncertainty analytics (especially those involve external information, intensive labour and personal intuition) has become less efficient and even invalid. On this basis, this thesis aims to propose efficient methods based on data mining techniques, which are less dependent on external resources, for the uncertainty evaluation, modelling and quality assurance in selected SBD types. Special attention has been paid to spatial vector data, trajectory data and spatial time-series data, which are amongst the most representative SBD types and significant in practical applications.

During spatial data production, quality assessment and control (QAC) is the primary process that controls data uncertainty and reliability. Traditionally, reference data are required during QAC for direct comparison to discover errors. However, in

the context of SBD, complete and accurate reference data are always unavailable for many reasons, such as the vast area of coverage and the administrative barrier of administrative divisions. In such a situation, developing reference-reduced or reference-free methods for QAC is necessary and promising. Therefore, this thesis started with a reference-free method to locate potential errors in multilayer vector data, which are the most representative data structure in practice. Spatial relationship complexity was adopted as an indicator for the identification of potential errors. The linkage between spatial relationship complexity and errors was initially discussed. A contribution function based on distance measurement was introduced to estimate the contribution of each vector layer, and the results were further taken as input into an entropy-based indicator to obtain the overall complexity measurement. On the basis of experiments on simulated and real-life datasets, the proposed approach outperformed state-of-the-art methods in providing realistic complexity measurements, and the resultant complexity map could provide useful information to facilitate manual inspection during QAC for large-scale vector data.

To extend our idea to a single vector layer, another reference-free method was proposed to identify potential classification errors in land use/land cover (LULC) data. In this method, land patches belonging to the same land class were assumed to present similar spectral-spatial features. In view of the influence of production scale in feature extraction, an adaptive segmentation strategy based on local variance index was designed to obtain homogenous segments. A clustering operation was further applied to the extracted features to distinguish outliers conservatively. Finally, an entropy-based indicator was developed to measure the likelihood of a land patch to be erroneously classified based on the clustering results. Experiments showed that the proposed method is superior amongst other state-of-the-art methods in terms of high accuracy.

During the QAC for traditional classification data, the influence from data production specification has always been neglected; however, this may be inapplicable

to SBD due to its vast data volume. For this reason, this study proposed an evaluation method, specifically for the uncertainty caused by the minimum mapping unit in LULC data. An assumption was initially made on the skewed distribution of land patch sizes and validated on open data. The optimal skewed distribution was determined through curve fitting technique. Thus, the omission errors could be evaluated based on the fitting results. The resultant omission errors were further used to estimate the commission errors by considering the conversion between land classes based on the statistics of their adjacency. Finally, a confusion matrix could be obtained to evaluate the overall classification accuracy. Experiments on real-life land cover dataset showed that the proposed method could accurately estimate the classification uncertainty for most land classes.

Trajectory and spatial time-series data are two of the most representative SBD commonly used in current studies on human behaviour, mobility and transportation. Therefore, specific efforts have been made to address prominent uncertainty issues of the two data types. For trajectory data, this study focused on modelling the uncertainty caused by sampling and measurement errors. This issue was selected because an effective uncertainty model is critical for many applications on trajectories, such as spatial query and visualisation. To reduce redundant uncertain regions in state-of-the-art models, an adaptive error ellipse (AEE) model was established, in which the optimal size for an error ellipse was obtained based on the Minkowski distance metric through mining the intrinsic characteristics in the trajectory data. A broad AEE model was further developed to include measurement error during the model construction, and an ellipse formulation was deduced to avoid intensive computation of the theoretical model and enhance practical applicability. Experiments on five real-life datasets showed that in comparison with the state-of-the-art methods, the proposed models could significantly narrow the uncertain ellipses while retaining a comparative accuracy. A case study on trajectory similarity analysis was further conducted to exemplify the practical advantages of the proposed models compared with the state-

of-the-art methods.

Lastly, this study discussed the uncertainty in spatial time series, and special attention was provided to evaluate its predictability. To reduce the inference from the randomisation in human behaviour on predictability evaluation, a novel evaluation method was proposed on the basis of entropy indexes and time series decomposition technique. Experiments were conducted on a real-life metro ridership dataset to validate the effectiveness of the proposed method. Results showed that the proposed indicators could reflect the evaluation values with higher correlation with the real predictability results than the traditional indicators. To further demonstrate its usefulness, an uncertainty-based loss function was implemented using the predictability measurements and was applied to the classical long short-term memory model for validation. Experiments showed that the proposed loss function could significantly improve prediction accuracy. The proposed method is theoretically extensible to other SBD in the form of time series.

Uncertainty will always be a major scope in future studies of SBD. The adoption of data mining technique may improve the execution efficiency of uncertainty analytics and data quality assurance and further enhance the reliability of related applications in a broad sense.

Publications arising from the thesis

[1] Shi, W., Chen, P., Zhang, X. 2017. Reliability Analysis in Geographical Conditions Monitoring. *Acta Geodaetica et Cartographica Sinica*,46(10) :1620-1626. (DOI: 10.11947/j.AGCS.2017.20170377)

[2] Chen, P., Shi, W., Kou, R., and Wan, Y., 2018. A quantitative investigation of the uncertainty associated with mapping scale in the production of land-cover/land-use data. *International Journal of Remote Sensing*, 39 (23), 8798–8817. (DOI: 10.1080/01431161.2018.1492179)

[3] Chen, P. and Shi, W., 2018. Measuring the Spatial Relationship Information of Multi-Layered Vector Data. *ISPRS International Journal of Geo-Information*, 7 (3), 88. (DOI: 10.3390/ijgi7030088)

[4] Chen, P., Shi, W., and Kou, R., 2019. Reference-Free Measurement of the Classification Reliability of Vector-Based Land Cover Mapping. *IEEE Geoscience and Remote Sensing Letters*, 16 (7), 1090-1094. (DOI: 10.1109/LGRS.2019.2893602)

[5] Chen, P., Shi, W., Zhou, X., Liu, Z., & Fu, X., 2019. STLP-GSM: a method to predict future locations of individuals based on geotagged social media data, *International Journal of Geographical Information Science*, 33 (12), 2337-2362. (DOI: 10.1080/13658816.2019.1630630)

[6] Shi, W., Chen, P., Shen, X., & Liu, J., 2020. An adaptive approach for modelling the movement uncertainty in trajectory data based on the concept of error ellipses, *International Journal of Geographical Information Science*. (Under review)

Acknowledgments

The Ph. D life was like swimming in a crowed and muddy river, sometimes making me suffocated and exhausted with endless stress. But thanks to the past, it is that desperate but beautiful experience allows me to feel the unprecedented ecstasy at this moment.

Thanks to my supervisor Prof. Wenzhong SHI, for his encouragement and guidance on my research. Without his support, I may not have the chance to study in Hong Kong and complete this dissertation. His strict self-requirement sets me an excellent example of what a distinguished scholar should be and will benefit my entire academic career.

I would like to thank the academic staffs from both Wuhan University and Hong Kong Polytechnic University for their support during my Ph. D study. My sincere gratitude goes to Prof. Alfred STEIN, Prof. Qingming ZHAN, Prof. Hao WU, Prof. Qingfeng GUAN, Prof. Penglin ZHANG, Prof. Changhui YU, Prof. Jiangping CHEN, Prof. Tao JIA, Prof. Wallace LAI. Thanks for their instructive comments and suggestions to improve my work.

I am also grateful to my colleagues and friends, sincere thanks go to Dr. Wan Yiliang, Ms. Kou Rong, Dr. Gao Lipeng, Dr. Hao Ming, Dr. Zhang Anshu, Dr. Zhao Yuanlin, Dr. Zhang Xiaokang, Dr. Li Zhenxuan, Mr. Zhang Min, Mr. Xiang Haodong, Mr. Liu Zhewei, Ms. Zhou Xiaolin, Mr. Nie Mingyan, Mr. Fu Xuandi, Mr. Chen Shanxiong, Mr. Wang Rui, Ms. Pan Hong, Ms. Chen Ling for their consideration and genuine advises, which got me through the hardest time in the past four years.

I especially thank my parents for their continuous support and love.

Table of contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Background and motivation

1.1.1 Emerging spatial big data (SBD)

We are living in an era of big data (Lohr 2012, McAfee et al. 2012). Since the concept was proposed in 2005, big data has become the dominant scientific paradigm in various disciplines (Kitchin 2014, Jose 2014, Walker 2014). Indeed, with rapidly emerging modern technologies in a high-speed network environment, the unique power of big data has put a worldwide wave of innovations to the entire academic and industrial community (Chen et al. 2012, Gobble 2013, Lee et al. 2014, Wolfert et al. 2017).

SBD, also known as geospatial big data, is the nomenclature when handling big data related to geographic space in geospatial information science (GISci; Li et al. 2016). SBD has inherited the 4V dimensions from big data, namely, volume, variety, velocity and veracity, and it is facing the problem from its spatial-oriented characteristic, such as visualisation, variability and value (Yu et al. 2014, Pavlyuk 2017). Generally, SBD can be categorised as 'earth observation data', such as satellite imagery and land cover data, and 'human activity data", such as vehicle tracks and geotagged social media data (Shi et al. 2018).

With the booming development of sensor techniques, network services and high-performance computing technologies, SBD has brought unprecedented opportunities and reflashes GISci from developing rudimentary tools to modern science for understanding and planning our world (Gerhardt et al. 2012, Lee and Kang 2015, Li et al. 2016). At present, SBD has been widely adopted from the traditional application area in earth science (Anselin et al. 2006), agriculture (Kamilaris et al. 2017) and mining industry (He et al. 2017) to modern applications such as urban planning (Batty 2013), environmental science (Li et al. 2016), archaeology (Casana 2014), sociology

(Leszczynski 2015) and public health (Lopez et al. 2014).

The popularity of SBD is believed to keep increasing. According to an open report prepared by Oxera (2013), the compound annual growth rate for the GIS industry can be safely estimated as 15%–20%, thereby extending the scale and impact of SBD. In addition, with the continuous development of location-based devices, such as mobile phones and GPS trackers, SBD will cover more areas of the world with fine location granularity (Dalton and Thatcher 2015). As a result, SBD has drawn considerable concerns from the academic community, general public and governments (Jones 2014).

1.1.2 Implication of SBD uncertainty

In GISci, uncertainty has been widely acknowledged as one of the fundamental elements of spatial data (Goodchild and Gopal 1989, Shi 2009, Shi et al. 2018). The concept of uncertainty represents the inaccuracy, fuzziness and randomisation of geospatial data (Shi 2009) and conveys information about data quality (Guptill 2017). Uncertainty issues arise in the entire process of SBD analytics; such issues include (1) uncertainties in real-world geographic entities and relations, (2) uncertainties in data capture, (3) uncertainties from human cognition, data preprocessing and data abstraction and (4) uncertainties in knowledge extraction (Shi et al. 2018). Once generated, uncertainties will propagate in all the aforementioned stages and significantly affect the reliability of extracted knowledge (Shi 2009).

Uncertain SBD and unreliable SBD analytic results can cause severe losses. According to International Business Machines (IBM), in the US, data quality issues, which are only part of uncertainties, lead to a loss of 3.1 trillion USD per year (IBM Big Data and Analytics Hub), which equals to 1/6 of yearly GDP in the US or 1/4 of yearly GDP in China. SBD uncertainty has also received increasing attention from the public and government. On the basis of the result of a recent survey about the future of GIS, the issue on data accuracy, which also refers to spatial uncertainty, draws the

most concerns from 32% respondents (De Milliano 2017). In September 2017, the China State Council published 'Suggestions on Deepening the Reform Environmental Monitoring and Improving the Environmental Monitoring Data quality', which is the first time in history that the State Council directly issued an official file that emphasises the critical role of spatial data quality (http://www.gov.cn/zhengce/).

The study on SBD uncertainty is of high theoretical and practical values. Researchers have long realised the implication of SBD uncertainty. Shi (2009) noted that spatial data quality control and uncertainty modelling is 'one fundamental area of GISci'. Wang (2011) noted that the development of uncertainty theory remarkably contributes to data quality control and guarantees the reliability of the application of geospatial data. Kwan (2016) noted that uncertainty is 'an essential element in the geographic knowledge production'. Schwanen (2018) claimed that 'greater attention should be paid to challenging forms of uncertainty, for which better data or analysis techniques are no panacea'. Shi et al. (2018) noted that 'the value of SBD relies on uncertainty handling and the reliability of extracted knowledge'. Goodchild (2018), in his latest article on the 'Annals of GIS', noted that many traditional methods for modelling uncertainty are thought to be highly complex and difficult to apply in practice. He also claimed that the development of uncertainty theory and methods would be an essential component that helps reinvent the GIS; 'with today's computing power and visualisation capabilities, the result might be very different'.

Given the significance of SBD uncertainty and the variety of SBD, this thesis focused on developing sound methods for uncertainty analysis and data quality assurance in representative SBD.

1.2 State of the art on SBD uncertainty analysis

1.2.1 Development of spatial uncertainty

Uncertainty generally exists in the objective world (Shi 2009). The concept of uncertainty principle was first introduced by German physicist Werner Heisenberg in

his pioneering work in quantum mechanics (Heisenberg 1925), which stated that one can never simultaneously obtain the exact position and momentum of some particles. Once this controversial statement had been validated, uncertainty principle was further expanded into general popular culture and other scientific domains, such as mathematics (Gödel 1931), thermodynamics (Prigogine et al. 1961) and economics (Arrow 1978).

Uncertainty has always played an essential role in spatial science. A famous example can be traced back to the famous question proposed by Polish mathematician Benoît B. Mandelbrot about the length of the British coastline (Figure 1.1; Mandelbrot 1967). The answer to this question is not unique, and this uncertainty involves many factors, such as personal perception, measurement methods and even measuring time. To reduce the influence of measurement error during data processing and obtain accurate spatial data, early efforts had been made on developing surveying error theory (Li 1986, Wang 1990).

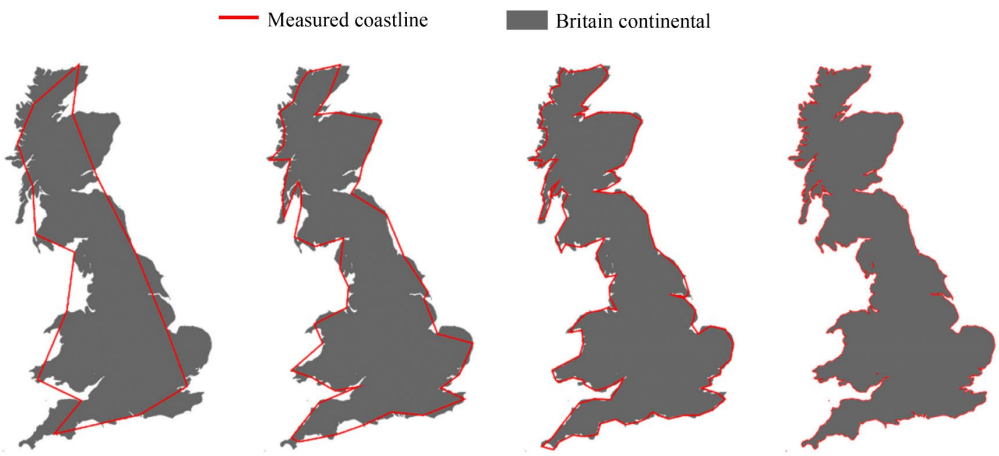


Figure 1.1 Britain coastline.
(https://commons.wikimedia.org/wiki/File:Britain_Fractal_Coastline.gif)

Representing the objective world using digital features, such as spatial points, lines and polygons, has entered the digital mapping era and has become mainstream in related studies. Conversely, the development of remote sensing techniques has

greatly enhanced the efficiency of conducting large-scale earth observation. As a result, the spatial uncertainty theory has received unprecedented attention and has rapidly developed. A series of theoretical and methodological research has been conducted emphasising on different aspects of spatial uncertainty, including spatial uncertainty models, spatial analysis uncertainty theory and spatial data quality control (Shi 1998, 2009, Fisher 1999, Shi and Liu 2000, Zhang and Goodchild 2002, Goodchild *et al.* 2003, Wang *et al.* 2005, Tong and Shi 2010). Specifically, Shi (2009) conducted comprehensive research elaborating the uncertainty theory system of spatial data and spatial analysis. He also pointed out the prospects of uncertainty theory in spatial data quality control and data mining and demonstrated a GIS data quality information system based on Webservice.

The study on spatial uncertainty has also served as a foundation for improving the reliability of spatial data and spatial analysis. Reliability is the reflection of spatial uncertainty at the application end (Shi et al. 2017). At present, given that the research on the quality and uncertainty of earth observation data is relatively mature, studies on spatial reliability almost focus on traditional vector data and raster data. Shi et al. (2012) discussed the theory in reliable spatial analysis and proposed quantitative measures for evaluating the reliability of spatial data. The research findings were further applied to a national surveying project and reported to enhance the data reliability supporting further decision making (Shi et al. 2015, 2017).

In the era of SBD, the uncertainty problem becomes more severe. Although SBD has enabled an extensive range of data analysis and practical applications, its uncertainty involves a wider variety, and the propagation mechanism becomes even more complicated. On the one hand, data production is no longer the exclusive work for only authorised departments. With the emergence of collaborative projects, such as the OpenStreetMap (OSM) project (Haklay and Weber 2008), public users can fully utilise their personal experience and knowledge and easily participate in the production of worldwide editable maps. Given the relatively loose production

specifications and the participation of nonprofessional users, these volunteered geographic information (VGI) data always possess uncertainty, and their quality is widely in question (Sui et al. 2012). Furthermore, as comprehensive and accurate reference data are always unavailable for these VGI, the traditional quality assurance methods become impossible (Fonte et al. 2017). To address these problems, researchers have assessed VGI data using crowdsourcing, social, geographic and even data mining approaches (Goodchild and Li 2012, Senaratne et al. 2017). Amongst these approaches, data mining methods have shown their prominent performance in VGI assessment as they examine the reliability of VGI data solely through internal or external consistency and require no reference with hand (Hashemi and Abbaspour 2015, Fonte et al. 2017). However, a systematic framework for VGI uncertainty is far from being established (Senaratne et al. 2017). On the other hand, the development of the mobile communication network and sensor technologies allow real-time and wide-range data collection for human activities. Study on the uncertainty of these human activity data involves a wide range of topics, such as geographic contextual uncertainty (Kwan et al. 2019), uncertain data mining (Shi, Zhang, and Webb 2018, Liu et al. 2019), data biases evaluation (Zhao et al. 2016, 2019), trajectory uncertainty modelling (Jeung et al. 2014), location uncertainty (Wan et al. 2017) and uncertainty visualisation (Huang and Wong 2015). However, at present, given the various characteristics of human activity data, a systematic framework for its uncertainty research is unavailable.

1.2.2 Uncertainty of representative SBD types: Principles and problems

Quality assurance is still amongst the most critical tasks in the uncertainty analysis of specific SBD, such as large-scale vector data and land classification maps. Traditional quality assessment and control (QAC), during which reference data are commonly required (ISO 2013), is still useful in some instances. However, in the context of SBD, due to rapid data updates, vast data volume and nonexistence of alternative

official/professional sources for many crowdsourced data types, performing a large-scale manual inspection is impossible, and reference data with adequate size and quality can be unavailable (Foody and Boyd 2013, Joshi et al. 2016, Goodchild 2018). This problem becomes even more significant for some production projects, such as global mapping and surveying, in which data quality is taken as the core and the primary indicator for project acceptance (Chen et al. 2015). Many solutions have been suggested to handle this problem. One common approach is to reduce the workload of QAC through optimal strategies. For example, a two-rank sampling plan was designed for the quality acceptance of geospatial data, which can achieve high acceptance quality level with a lower sampling rate than commonly used standards (Tong et al. 2011). External information has also been adopted to reduce the reliance on reference data. Taking the quality assessment of VGI as an example, which has been briefly discussed in the last section, a batch of indicators have been proposed based on the correlation between data quality and contributors' status, demographic and socioeconomic status or even the information in metadata (Ciepluch et al. 2010, Zielstra and Zipf 2010, Antoniou and Skopeliti 2015, Mullen et al. 2015); another batch of methods solely examine VGI dataset and aim to assess data quality by purely using data mining methods, such as outlier detection, cluster analysis and regression analysis (Foody 2012, Barron et al. 2014, Senaratne et al. 2017). These methods have provided great potential in solving quality-related issues for other SBD. However, given the heterogeneity of SBD and limitations in existing methods, it is far from building a comprehensive theory system and robust methodology for the QAC in SBD.

Furthermore, some uncertainties that are traditionally considered negligible may cause a significant influence in the context of SBD. This phenomenon is prominent in the uncertainty brought by data production specifications. Conventionally, rules in specifications, such as the definition of the classification system and minimum mapping unit (MMU), are considered rational, and their potential influences on final products are ignored. This condition is reasonable for traditional spatial data

production because, on the one hand, the potential uncertainty may be insignificant due to relatively small data coverage, and on the other hand, these rules are technically applicable as they are determined by professionals after fully considering the characteristics of the target region (Jansen and Gregorio 2000, ISO 2013). However, evidence can be found in current studies that these uncertainties may not be negligible, especially for large-scale production projects with high heterogeneity in different regions (Saura 2002, Pascual-Hortal et al. 2007, Rutchey et al. 2009, Kelly et al. 2011). Leaving such uncertainties unsolved will result in biased statistics and analysis results. Nevertheless, few efforts have been made in this direction.

In comparison with earth observation data, the uncertainty in new-fashioned human activity data is a cutting-edge direction in GISci community. As discussed in the last section, human activity data involve many data types and various uncertainty issues. Particularly, trajectory big data are amongst the most representative and prominent data types because they have been broadly adopted in the intelligent transportation system, urban planning, mobility analysis and behaviour pattern mining (Zheng 2015). Except for the temporal dimension, a trajectory is generally similar to the points and lines in traditional vector data; thus, some traditional uncertainty theories and methodologies for vector data can be adopted for trajectory data. At present, studies on the uncertainty in trajectory data mainly focus on uncertainty modelling. Different models, such as cylinder and beads models (Trajcevski et al. 2004, 2010, Kuijpers and Othman 2006), have been proposed to measure the position uncertainty of moving objects over time and benefit a wide range of practical applications, such as uncertainty-based spatial query and visualisation (Niedermayer et al. 2013a, Huang and Wong 2015, Furtado et al. 2018). However, these models always involve empirical or intuitive parameter setting, whereas trajectory uncertainty can be various and related to many factors, such as moving speed and distance. This problem leads to less robustness for these models in practical use, which decreases the reliability of their further application.

In addition to trajectory data, spatial time series is another representative data type in SBD. Time-series data can capture the dynamics of a given statistical indicator over time. Forecasting and simulation are the main functions for time-series big data in many disciplines, including GISci (Davis and Palumbo 2001, Gutiérrez et al. 2011, Brockwell and Davis 2016). Different from other SBD data, the uncertainty of spatial time-series data is always embodied in its predictability rather than data accuracy. Using unpredictable time series may result in unrealistic prediction models and forecasting results (Clements and Hendry 2000). Entropy is commonly used in estimating the predictability of time series, and many indicators have been proposed, such as approximate entropy (ApEn), sample entropy (SampEn), fuzzy entropy and multiscale entropy (Pincus 1991, Richman and Moorman 2000, Costa et al. 2002, Borowska 2015). However, considering the underlying randomisation in human behaviour and potential errors during data collection, human behaviour time series can be compared to a continuous measurement with random measurement errors. These random errors can introduce complicated influence on those predictability indicators that treat the time series as a whole, which results in biased predictability measurements. Therefore, new methods that can reduce the negative influence from the random parts in time-series data should be developed to obtain realistic predictability measurements.

1.2.3 Applying data mining methods to handle SBD uncertainty

Although voluminous SBD can be obtained, we know less about the properties of data and can hardly conduct efficient analysis solely based on speculation and gut feeling. To address this problem, data mining, which serves as a powerful tool for extracting knowledge from big data (Fan and Bifet 2013, Wu et al. 2013), exhibit great potential. Data mining aims to achieve the automatic or semi-automatic analysis of large datasets (Klemettinen et al. 1997, Hand 2006). Through data mining, previously unknown and unexpected data patterns can be extracted. These patterns can reflect inherent

characteristics of the dataset and provide useful information supporting for subsequent data analysis. Over the past decades, data mining has made many remarkable achievements in the field of big data analysis, such as the surprising applications in the stock market (Preis et al. 2013) and flu (Butler 2008).

The potential of data mining in SBD uncertainty analysis has long been proved in literature. For example, in the comprehensive review of VGI quality assessment, Senaratne et al. (2017) claimed data mining as one of the major assessment approaches considering its significant capability in discovering patterns purely from data with no requirement on geographic laws and knowledge. This work also summarised the conventional data mining approaches in the literature, including clustering analysis, latent class analysis, correlation statistics, outlier detection, heuristic metrics and fuzzy logic. For human activity data, data mining also plays a vital role in optimisation and self-adaptation issues. Zheng (2015) argued that data mining is one of the major techniques for handling trajectory uncertainty. However, the data mining process can also introduce uncertainties (Shi et al. 2003, Zhang et al. 2016). Thus, people need to be careful when using data mining in addressing uncertainty issues, and in some cases, a certain amount of human involvement will help improve the reliability of mining results.

1.3 Scope and objectives of this study

This study aims to improve the current QAC process for SBD to facilitate the reduction of reliance on traditional reference data. This study also aims to establish sound models and indicators to represent and estimate the uncertainty in representative human activity data. Particularly, this study has three objectives:

- To develop reference-free methods to enhance the QAC process for earth observation products;
- To propose sound methods for adaptively modelling the position uncertainty in trajectory big data;

- To design sound indicators for evaluating the uncertainty in spatial time-series big data.

Particularly, in addressing earth observation products, the scope of this thesis has been limited to spatial vector data because the QAC process of spatial vector model is more representative than that of spatial raster model in GISci. Spatial vector model is also one of the significant geospatial products in surveying and mapping projects. Thus, the study on QAC process and uncertainty is significant and promising.

1.4 Structure of the dissertation

This dissertation consists of seven chapters. Chapters 2–4 introduce a series of reference-free methods for the QAC of earth observation data. Chapters 5 and 6 demonstrate novel methods for uncertainty modelling and evaluation for trajectory big data and spatial time-series big data. Chapter 7 concludes this thesis. The structure and content of the seven chapters are summarised as follows.

In Chapter 2, an entropy-based method was proposed to locate the potential quality issues in multilayer vector data. The proposed method assumes the linkage between spatial relationship complexity and quality issues and capture regions where errors tend to appear based on complexity measurement. An entropy-based indicator for measuring the complexity was proposed by considering the effect from a single vector layer and the combined effects from multiple layers. The effectiveness of the proposed method was demonstrated through the experiments on real-life and simulated datasets.

In Chapter 3, a reference-free method was further established to discover and identify potential classification errors in land use/land cover (LULC) data. In this method, land patches were initially divided into homogenous segments at an optimal scale, and spectral and textural features were further extracted. Then, an adaptive clustering process was conducted to identify the clustered segments and outliers. On the basis of clustering results, an entropy-based measure was proposed to describe the

likelihood of a land patch to be wrongly classified. The results on two real-life datasets and the comparison with state-of-the-art methods demonstrated the efficiency of the proposed method.

In Chapter 4, a reference-free approach was proposed to evaluate the uncertainty brought by the MMU in LULC mapping. In this approach, the patch size of a land class was assumed to obey a positively skewed distribution. On the basis of this assumption, the proposed approach estimated the omission error through a curve-fitting method with multiple customised transformation functions. Then, the commission error was evaluated by analysing the adjacency relationship between land patches and calculating the conversion probabilities amongst land classes. Finally, on the basis of the estimated omission and commission errors, a confusion matrix could be constructed, and the classification accuracy could be estimated without reference data. Tests on real-life and simulated datasets demonstrated the effectiveness of the proposed approaches.

In Chapter 5, two models were proposed to represent the uncertainty in trajectory big data. The first model was constructed to reduce the redundant area of error ellipses in state-of-the-art methods. To achieve this objective, an adaptive distance metric was obtained on the basis of the underlying characteristics of each trajectory to generate the most appropriate error ellipses. The second model improves the first model by considering the measurement error on trajectory points and generating an ellipse-like region to represent the trajectory uncertainty. To enhance computational efficiency, these regions were approximated by standard ellipses, of which the detailed derivation process was discussed. Experiments were conducted on five real-life datasets to validate the effectiveness of the proposed models.

In Chapter 6, an entropy-based indicator was proposed to measure the predictability of spatial time-series data. Four widely used prediction models were adopted to approximate the real predictability, and the results were further used to validate the proposed indicator. Experimental results indicated a high and stable

correlation between the measured and the approximated predictability. Furthermore, to demonstrate the usefulness of the predictability indicator, it was implemented into an uncertainty-based loss function to improve the performance of traditional neural network models in prediction tasks. The experimental results showed that the proposed loss function could bring an average improvement of 4.19% on prediction accuracy for the case data.

In Chapter 7, a summary was presented. In short, this study proposed a series of reference-free methods, especially for the QAC process of SBD, established novel models for the uncertainty in trajectory big data and proposed a robust evaluation method for uncertainty/predictability in spatial time-series data. Advantages, limitations and prospects of this study were also discussed.

Figure 1.2 shows the logical structure of this thesis.

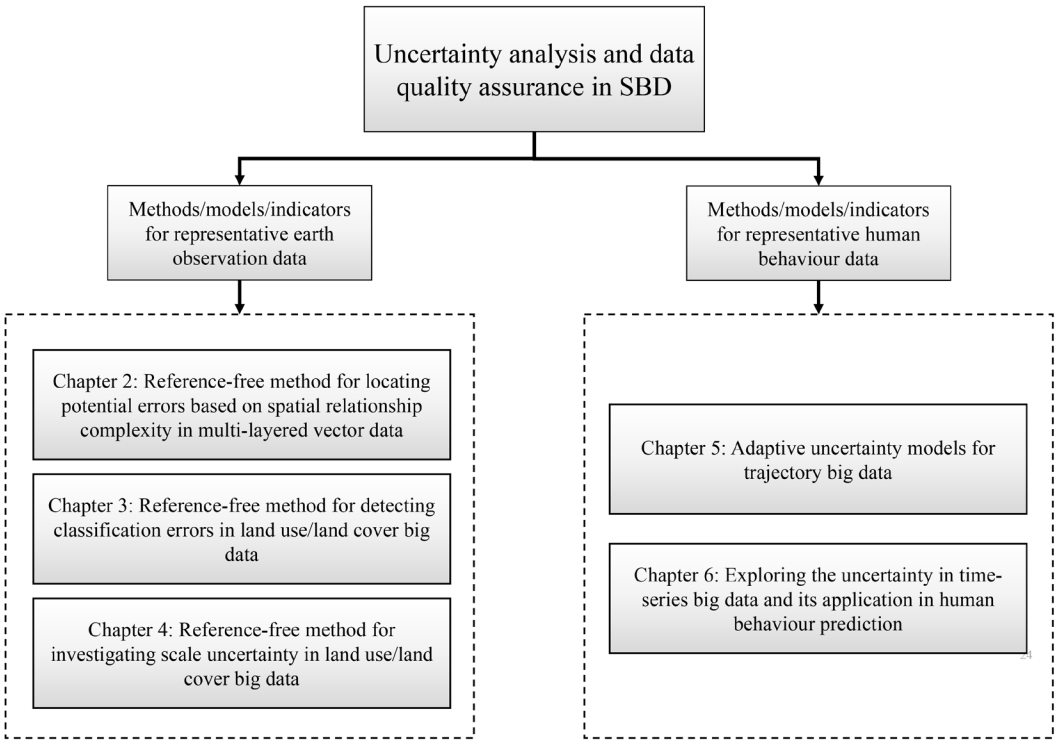


Figure 1.2 Thesis structure.

Chapter 2 Reference-free method for locating potential errors based on spatial relationship complexity in multilayer vector data

Spatial vector model has provided a useful framework for encoding spatial relationships and manipulating spatial data. Spatial relationship plays an important role in the quality assurance for spatial vector data. On the one hand, spatial relationship is highly related to logical inconsistency, which is acknowledged as one of the essential elements for spatial data quality (ISO 2013). On the other hand, in complicated spatial relationships, many spatial objects may be involved, and the risk of inconsistency amongst these objects will increase, which result in quality issues to the dataset. In that sense, the spatial relationship complexity can be an indirect indicator for identifying potential quality issues. A reference-free evaluation method is introduced in this chapter to estimate the complexity of spatial relationship and locate the regions where quality issues tend to occur in multilayer vector data.

2.1 Overview of spatial relationship complexity

Generally, spatial relationship includes topological, distance and directional relationships (Egenhofer and Sharma 1993a). Sufficient information about spatial relationships can boost the understanding and reasoning of spatial situations (Frank 1992, Li 2007).

Study on the complexity of spatial relationships has a long history, in which entropy always serves as a powerful tool for complexity measurement. The concept of entropy was proposed by Shannon (1948), which later on formed the foundation of information theory (Shannon and Weaver 1949). In information theory, entropy is widely used to measure the disorder, diversity and complexity of a system (Jost 2006). Formally, given a categoric variable T with n classes, Shannon's entropy (ENT) can be expressed as

$$H(T) = -\sum_{i=1}^{n} p_i \log(p_i), \tag{2.1}$$

where p_i denotes the probability for the variable belonging to the ith category.

Thus far, entropy has been widely applied in GISci, such as scale effects, landscape analysis and spatial heterogeneity (Chen and Sun 2014, Leibovici et al. 2014, Gao et al. 2017). The application of entropy on measuring spatial information or spatial complexity can be traced back to the pioneering work of Sukhov (1967), who adopted entropy to measure the diversity of objects on a map. Thereafter, the idea was improved and extended to other aspects of spatial information, such as topological, geometric and thematic information (Neumann 1994, Bjørke 1996, Li and Huang 2002).

However, subjects in the aforementioned methods are either a single-layer map or a composite layer with all the features integrated (Figure 2.1(a)), which makes them inappropriate for multilayer vector data. As shown in Figure 2.1(b), objects of multilayer vector data are stored in different layers in a multilayer structure according to the features of their counterparts in the real world. Although some scholars have suggested standalone evaluation for the spatial complexity of each layer (Wang et al. 2007), this condition can only expose the complexity generated in a single layer while overlooking the intervention from multiple layers on spatial relationships.

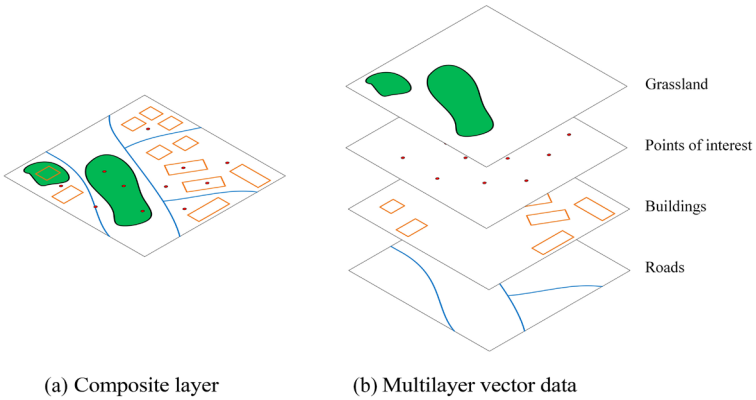


Figure 2.1 Examples of multilayer vector data and composite layer.

2.2 Complexity contribution from a single layer

2.2.1 Influence of a single spatial feature

A concept of contribution field is introduced in modelling the influence of spatial features on its surrounding to measure the complexity of multilayer vector data. According to the First Law of Geography, close objects tend to be more related than distant ones (Tobler 1970). Thus, on the basis of inverse distance weighting, which is commonly used in spatial interpolation (Baczkowski and Clark 1981, Bartier and Keller 1996), the contribution I of a specific grid cell x is defined as a decreasing function with respect to the shortest distance d to the closest spatial feature. In addition, a parameter λ is added as a control factor to limit the descending speed. Therefore, the contribution function I can be expressed as

$$I(x) = (d + \lambda)^{-1} \quad . \tag{2.2}$$

Curves of the contribution function with different values of λ are plotted in Figure 2.2. The descending speed of the contribution function increases (the curve becomes steeper) with the decrease in the value of the control factor λ. In practice, the control factor λ can be adjusted through visualisation to easily interpret complex regions.

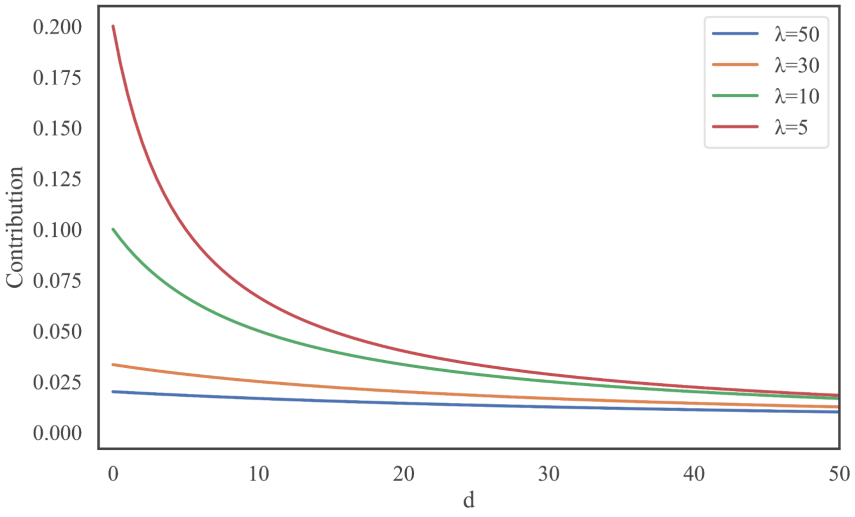
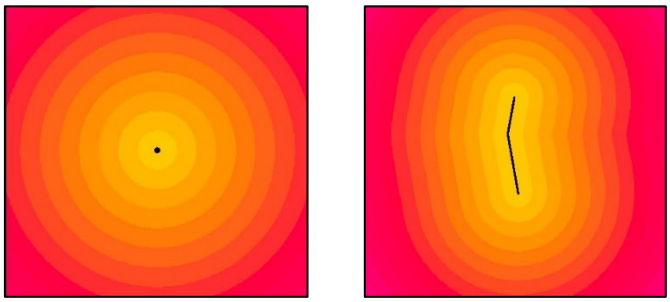


Figure 2.2 Contribution values with different parameters.

2.2.2 Contribution field for different types of feature

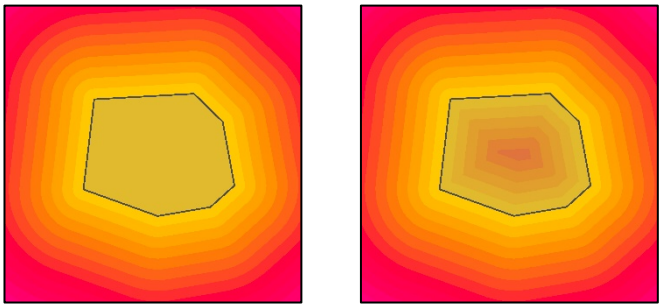
A contribution field is defined as the spatial distribution of the contribution of a spatial object on its surroundings. As shown in Figure 2.3, the contribution field of points (Figure 2.3(a)) and lines (Figure 2.3(b)) can be easily obtained as a radiation area. However, in comparison with points and lines, polygons involve more complicated spatial topological relationships (Chen et al. 2001). The boundary of a polygon is of essential importance in such relationships because of its critical role in distinguishing 'contains' and 'disjointed' cases as well as in other relationships, such as 'overlaps', 'meets' and 'equals' (Egenhofer and Sharma 1993b). Furthermore, distinguishing the importance of the interior and exterior of a polygon in contributing complex spatial relationships is difficult, especially when the attributes of features are unknown.



(a) (b)

Figure 2.3 (a) Contribution field for a point. (b) Contribution field for a line.

In this study, polygons are converted into lines before generating the contribution field. Examples for the contribution field of a polygon before and after the conversion are shown in Figure 2.4.



(a) (b)

Figure 2.4 (a) Contribution field for a polygon before conversion. (b) Contribution field for a polygon after conversion.

A contribution map should be initially defined to model the contribution of a single layer. In a contribution map, each grid cell should be assigned the sum of contribution values from all the features on a map. However, this operation is not computationally efficient because the distances between each grid cell and all features need to be computed. The calculation rule is modified with three steps for efficiency. Firstly, when we calculate the contribution value of a grid cell, buffers are generated for features, and only the features of which the buffer contains the grid cell are counted. The selection of buffer size will be discussed in Section 2.4. Secondly, we amplify the contribution from the nearest feature *N* times, where *N* is the number of counted features in the first step. Thirdly, the amplified contribution values are assigned to the corresponding grid cells, and the contribution map can be obtained. In this manner, the contribution map can be generated using three basic GIS analytics, namely, buffering, overlapping and distance analysis.

2.3 Combined complexity from multiple layers

2.3.1 Entropy-based measure for spatial relationship complexity

For a grid cell, let m be the number of layers and $I_j(x)$ be the contribution value on grid cell *x* based on the contribution map of the jth layer. Thus, a contribution set is defined as $S_x = \{I_1(x), I_2(x), \dots, I_m(x)\}$, which refers to the combination of contribution values from m layers. Then, the proportion of one contribution value in this contribution set can be calculated as follows:

$$p_j = \frac{I_j(x)}{\sum_{j=1}^{m} I_j(x)}. \tag{2.3}$$

The disorder of a contribution set is an essential element to describe the complexity at the given grid cell. The contribution value from a given layer on a grid cell reflects the influence of that layer on the spatial relationships in which the grid

cell is involved. In a situation where multiple layers contribute similarly to the grid cell, the spatial relationships can be taken as the results of the combined effects from these layers. In another situation where only a few layers contribute the most to the grid cell, the spatial relationships seem to be more straightforward because only these layers are involved, whereas the remaining layers are excluded. Therefore, the fundamental thought of our approach is that a contribution set has high complexity if it contains similar contribution values.

In addition, the intensity, such as the absolute contribution values, of the contribution set and the number of layers should be considered. For example, a set $\{0.3,0.3,0.3\}$ should have higher complexity than $\{0,0.3,0.3\}$ and less than $\{0.7,0.3,0.3\}$. Moreover, the complexity of the set $\{0.3,0.3,0.3,1\}$ should be larger than that of set $\{0.3,0.3,1\}$ as the number of layers increases. Therefore, the classical entropy is reformulated, and the complexity of contribution set X is defined as

$$C(S) = C(I_1, I_2, .., I_m) = -\varphi(S)\sum_{j=1}^{m} p_j \log(p_j), \tag{2.4}$$

where $\varphi(S)$ is a nonzero positive coefficient with respect to the intensity of the contribution set.

According to the above discussion, $C(S)$ should satisfy the following criteria.

Criterion 1: Increase with respect to I_j, mathematically, that is,

$$\frac{\partial H(S)}{\partial I_j} \geq 0 \qquad \forall I_j \in S. \tag{2.5}$$

Criterion 2: Ascend with the number of layers, that is,

$$C(I_1, I_2, .., I_m) \leq C(I_1, I_2, .., I_m, I') \qquad \forall I' \in [0,1]. \tag{2.6}$$

Criterion 1 implies that the derivative of $C(\mathrm{S})$ with respect to x_i for any $j = 1,2, \ldots, m$ should be above zero, that is,

$$\frac{\partial C(S)}{\partial I_j} = -\varphi_{I_j}(S)\sum_{j=1}^{m} p_j log(p_j) + \frac{1}{I_{sum}}\varphi(S)\sum_{j=1}^{m} p_j log(p_j) - \frac{1}{I_{sum}}\varphi(S) log(p_j) \geq 0, \quad (2.7)$$

where $I_{sum} = \sum_{j=1}^{m} I_j$, $\varphi_{I_j}(S)$ is the partial differential of $\varphi(S)$ with respect to the contribution variable I_j and p_j is the proportion of I_j in the given set S.

The general solution for (2.7) can be easily obtained as follows:

$$\varphi(S) = kI_{sum} = k\left(I_1 + I_2 \cdots + I_m\right), \quad (2.8)$$

where k is a positive constant. Equation (2.8) satisfies Criterion 2 when the constant k is set to 1. Therefore, the complexity $C(S)$ of a given contribution set can be written as

$$C(S) = -\sum_{j=1}^{m} I_j \log(\frac{I_j}{\sum_{j=1}^{m} I_j}). \quad (2.9)$$

According to Equation (2.2), it is easy to know that I reaches its maximum $\frac{1}{\lambda}$ when $d = 0$. According to the definition of $C(s)$ in Equation (2.4), when $I_j = \frac{1}{\lambda}$ for all $j = 1,2,\ldots,m$, $\varphi(S)$ and the rest entropy part simultaneously reach their maximum so that the upper bound of $C(s)$ can be obtained. In other words, when vector objects from each layer appear in the same grid cell, the complexity of this grid cell reaches the maximum, which can be expressed as follows:

$$C(S)_{\max} = C(I_1, I_2, ..., I_m|_{I_1 = I_2 = ... = I_m = 1/\lambda}) = \frac{m}{\lambda}\log(m). \quad (2.10)$$

Furthermore, considering all grid cells within the data extent Ψ, the total complexity C_{total} of a multilayer vector dataset can be expressed as

$$\mathrm{C}_{\mathrm{total}}(\psi) = \sum_{x \in \psi} C(S_x). \quad (2.11)$$

To compare the complexity of maps with different extents, the average complexity is defined as

$$C_{mean} = \frac{C_{total}(\psi)}{N_{\psi}}, \tag{2.12}$$

where N_{ψ} denotes the number of grid cells within the layer extent Ψ.

2.3.2 Overall workflow

On the basis of the above discussion, the overall workflow of the proposed method can be summarised as follows.

1) Set $j = 1$. Calculate the Euclidean distance (ED) to the closest source for each grid cell and the corresponding contribution value.

2) Generate a buffer layer for vector layer *j*. The buffer size α theoretically corresponds to the longest distance that the influence from a feature is not negligible. For a polygon layer, the buffering operation should be processed based on the corresponding line features that are converted from the boundary of the polygon.

3) Divide the extent Ψ of the layer into k areas $\{\psi_1, \psi_2, \ldots, \psi_k\}$ according to the overlapping amongst the buffer.

4) Let n_k be the number of buffers that cover the area ψ_k. For each grid cell *x* within an area ψ_k, amplify the contribution value $I(x)$: $I(x) \leftarrow I(x) * n_k$. Then, the contribution map of layer j is obtained.

5) $j \leftarrow j + 1$. Return to Step 3 until all the layers are traversed.

6) Compute the complexity based on Equation (2.9).

Three simulated datasets are used to demonstrate the overall workflow of our method. The ED is calculated based on a map unit equalling to the height of a grid cell. The parameter λ is set to 1 for the contribution function, and the buffer size is set to four times or the grid height. To understand the internal computation of the proposed method, the distance diagrams, contribution maps and the resultant complexity maps are shown in Figure 2.5.

As shown in Figure 2.5, the two layers in Set 1 only have one feature. In Set 2, lines appear at the same location as Set 1 while only the number of features increases.

Therefore, the distance diagrams are the same for Sets 1 and 2, whereas the contribution map becomes amplified as more features appear in both layers in Set 2. Intuitively, Set 2 should have higher complexity due to the increasing number of features. The final result supports our expectation in which the total complexity is 16.71 for Set 1 and 38.38 for Set 2. On the other hand, in Set 3, where the features move closer and even intersect in contrast with Set 2, the spatial relationships become more complicated, and the complexity is measured as the highest value, that is, 57.20. These examples demonstrate the computation process of the proposed method and improve the consistency between the measurements and subjective recognition of spatial relationship complexity.

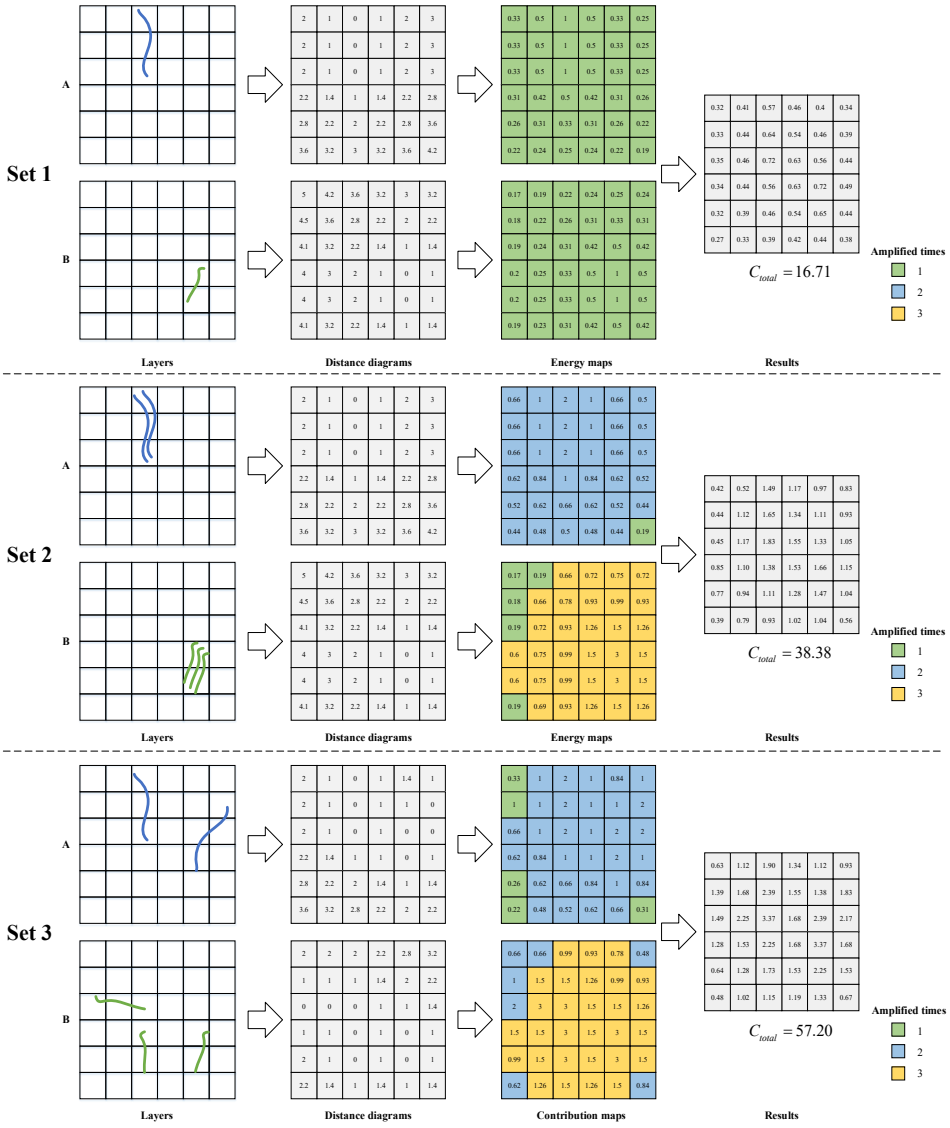


Figure 2.5 Three examples of the proposed method.

2.4 Experiments and analysis

2.4.1 Effects of the grid and buffer sizes

The selection of grid and buffer sizes are critical for the proposed method. The grid size determines the unit of the complexity measurement across the entire analysis process. The buffer size affects the degree of amplification while producing the contribution map. To conduct the sensitivity analysis on these two parameters, a simulated dataset is constructed, as shown in Figure 2.6, in which some essential spatial relationships, including ‘contains’, ‘intersects’ and ‘disjointed’, are designed.

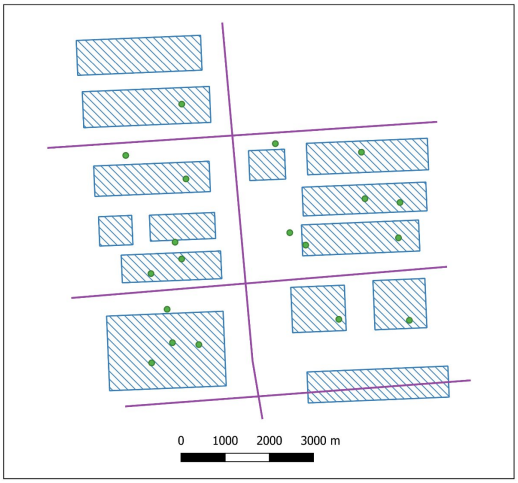


Figure 2.6 Simulated data. The extent of this simulated data is 10,000 × 10,000 m.

Multiple combinations of grid and buffer sizes are tested on the simulated data, and the fitting curves are plotted in Figure 2.7. The grid size has little effect on the complexity measurement when relatively small values are selected. However, when the grid size increases, a significant change is observed in the resultant complexity. This fact is reasonable because the data will be over-abstracted as the large grid size leads to the loss of considerable spatial information, such as shape information, during the computation of a contribution map. Therefore, the grid size should not be excessively large in practice. Conversely, the complexity considerably increases with the buffer size. This phenomenon can be expected because a larger buffer size

indicates that the contribution from more features will be counted, and thus the resultant contribution values will become higher.

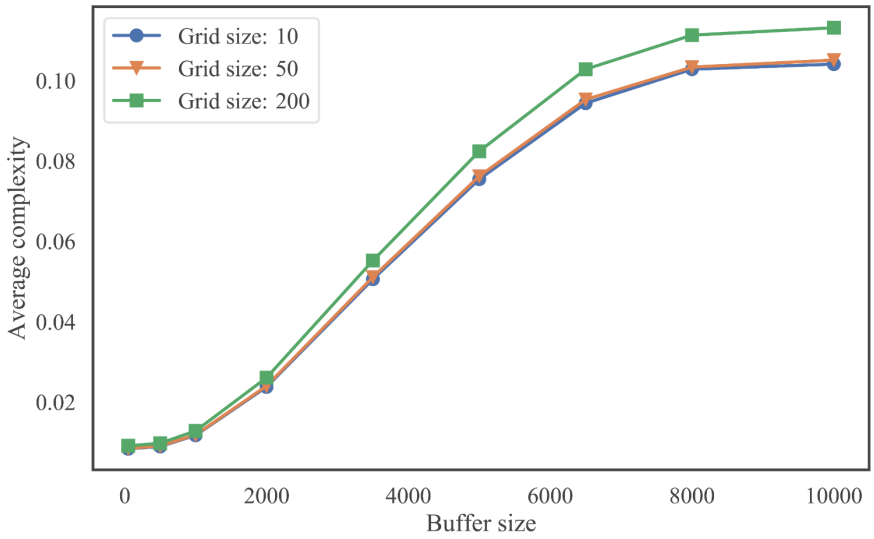


Figure 2.7 Plots of average complexity with various buffer and grid sizes.

Two points should be noted on the selection of grid and buffer sizes in practice. For grid size, an extremely small value will result in intensive computation during related spatial analysis, whereas a large value will lead to a high degree of abstraction, as discussed previously. Therefore, the value of the grid size is suggested to be no larger than the minimum bounding square of all spatial features, which can avoid information loss and save computation cost. Conversely, the selection of buffer size should rely on data. We suggest two rules for the selection of the appropriate buffer size in practice. Firstly, the buffer size should not be extremely large; otherwise, the computation will be costly during the buffering and overlapping analysis. Secondly, the average distance between features can be the default value for the buffer size. In this manner, the contribution from most features will remain, and the bias in the calculation of contribution maps will be reduced.

2.4.2 Experiments on real-life datasets

In this section, a set of real data is adopted for the validation of the proposed method. As shown in Figure 2.8, three regions were extracted from the OpenStreetMap (OSM;

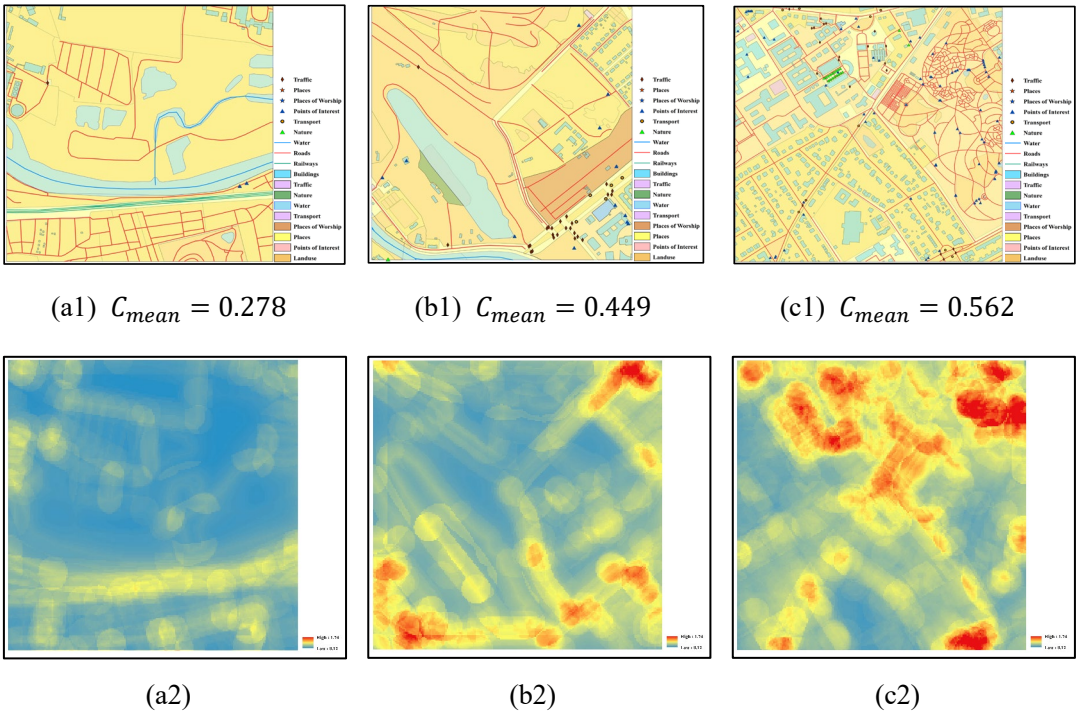
http://download.geofabrik.de/) data in Berlin, Germany. The experimental dataset comprises 18 layers, including 7 point layers, 3 line layers and 8 polygon layers. Table 2.1 presents the details of the OSM data in the three regions. The extent size is 1000×1000 m for each region, the buffer size is set to 50 m and the grid size is set to 5 m.

Table 2.1 Statistics of experimental data (/ means nonexistence).

Layer name	Type	Region (a1)		Region (b1)		Region (c1)	
		Feature number	Area/length	Feature number	Area/length	Feature number	Area/length
Traffic	Point	1	/	25	/	21	/
Places	Point	0	/	0	/	0	/
Places of worship	Point	0	/	0	/	0	/
Points of interest	Point	2	/	12	/	81	/
Transport	Point	0	/	5	/	9	/
Nature	Point	0	/	1	/	19	/
Water	Line	2	1,824	4	611	5	3,150
Roads	Line	107	12,782	33	10,896	320	27,605
Railways	Line	8	3,575	0	/	0	0
Buildings	Polygon	56	19,595	146	39,146	484	150,564
Traffic	Polygon	0	0	3	4,458	6	4,589
Nature	Polygon	0	0	1	8,740	0	0
Water	Polygon	11	132,068	5	109,353	0	0
Transport	Polygon	0	0	0	0	0	0
Places of worship	Polygon	0	0	1	73,922	0	0
Places	Polygon	2	1,000,000	3	1,000,000	2	1,000,000
Points of interest	Polygon	2	33,083	32	567,556	21	509,766
Land use	Polygon	21	888,516	31	1,177,776	36	513,450

As shown in Figure 2.8, the complexity of spatial relationships visually increases from left to right. The statistics of the data shown in Table 1 also support this inference

because the number of spatial features undergoes significant increases from region (a1) to (c1). On the basis of the complexity measurement of the proposed method, the average complexity of the three regions is computed as 0.278, 0.449 and 0.562. The results are consistent with our visual inference and seem reasonable. Conversely, the complexity maps of the spatial relationship are illustrated in Figure 2.8 (a2)–2.8(c2). These maps reveal the spatial distribution of the spatial relationship complexity, which can provide useful information for the location of potential quality issues.



(a1) $C_{mean} = 0.278$ (b1) $C_{mean} = 0.449$ (c1) $C_{mean} = 0.562$

(a2) (b2) (c2)

Figure 2.8 Complexity measurements of spatial relationships in three regions from the same dataset.

2.4.3 Comparison with state-of-the-art methods

On basis of the results of the last experiment, our approach is proven to provide an appropriate measurement of the spatial relationship complexity that is consistent with our observation and recognition. To further test the sensitivity of the proposed method in capturing the slight changes in the spatial relationships and its advantages compared with the classical methods proposed by Li and Huang (2002) and Liu et at (2012, 2013, 2013), a comparison test is conducted on three sets of simulated data. The multiple

layers are transformed into a single composite layer, as required in Li and Huang's method.

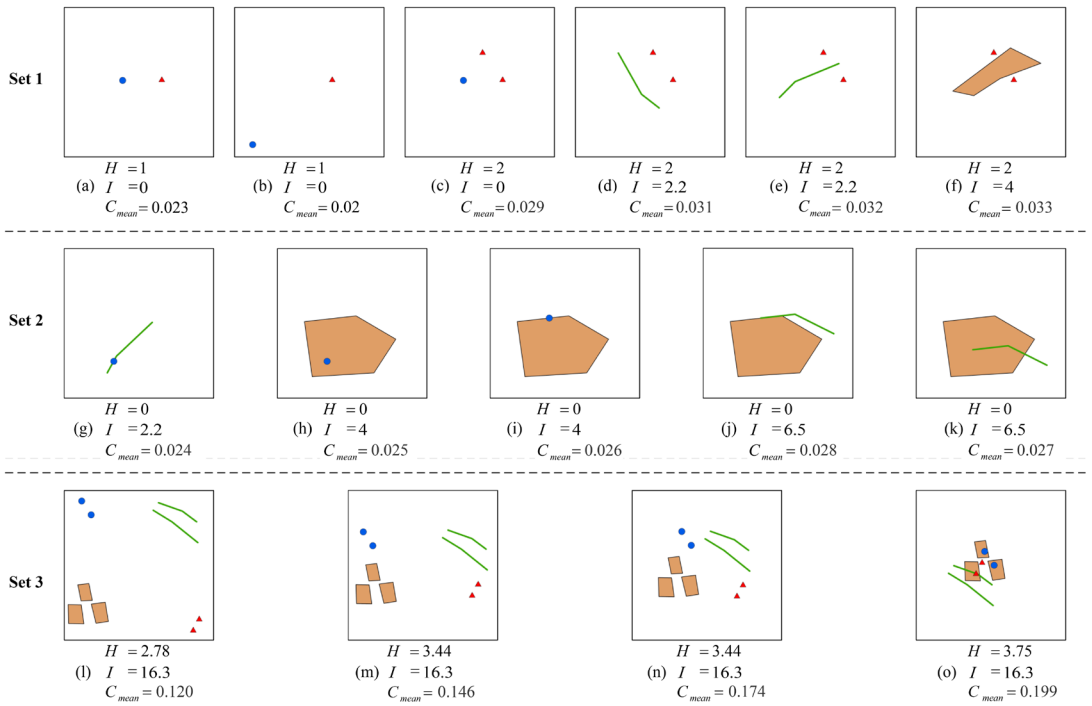


Figure 2.9 Comparison test on three sets of simulated data. H indicates the measurement based on Li and Huang's approach. I refers to the measurement of Liu's method. The extent is 100 × 100 m, the grid size is 1 m and the buffer size is 50 m. Layers are rendered in different colours.

The simulated data comprises four layers, including two point layers, one line layer and one polygon layer. As shown in Figure 2.9, Set 1 emphasises the 'disjointed' relationship, which is the most common situation in practice. On the basis of the measurements of Set 1, the proposed method can capture the small changes in the relative position ((a), (b) and (d), (e)); feature number ((a) and (c)); and feature type ((c), (e) and (f)). By contrast, the classical methods seem to be insensitive to those changes. The results of Set 2 indicate that the proposed method can capture topological changes, such as 'contains' ((g) and (h)), 'meets' ((i) and (j)) and 'overlaps' (k), whereas Li and Huang's method is inefficient in these cases with a measurement equalling to 0 and Liu's method can only capture the change of feature types in these cases. This is further improved by the general cases in Set 3, in which Liu's method

keeps to output the same measurement with different layouts. As the features become closer ((l) to (n)) and finally form a relatively complex layout (o), the complexity is expected to increase gradually. The proposed method shows better performance in reflecting the complexities of these four situations, which generates a larger coefficient of variation of 0.185, than the one of Li and Huang's method (i.e. 0.105). Three sets of data are designed for testing different perspectives of the proposed method. Set 1 aims to explore the sensitivity to the number, type of features and the distance between them. Set 2 validates the capability of distinguishing topological changes. Set 3 constructs general situations to test the robustness.

In fact, the objectives of the above methods are different: Liu and Li mainly aim to deal with the map generalization issue. Li's method treats vector dataset as a composite layer, using the Voronoi graph to measure the complexity, so it is theoretically inefficient to capture the varying relationship between intersected features as the Voronoi graph in such situation will not change. Liu's method processes layers solely and the models for points, lines and polygons are quite different. For this reason, Liu's method will not change when features from different layer get closer, because the method only adds the measurement of each layer together.

2.5 Summary

In this chapter, a novel method was proposed to locate potential quality issues in multilayer vector data. In this method, the complexity of spatial relationship was used as an indicator to represent the likelihood of the occurrence of quality issues. A contribution function was proposed to measure the complexity contribution from a single layer, and an entropy-based measure was further proposed to measure the overall complexity contributed by all layers. Finally, a complexity map was generated to display the distribution of complexity measurements.

In comparing with the state-of-the-art method on real-life and simulated datasets, it is difficult to say which method is the best since the purposes are quite different.

However, our method provides a relatively uniform framework to measure the topological complexity, being superior in addressing multilayered structure and proved more effective in capturing slight changes in data. From the view of application, this method is practically promising. Firstly, the complexity measurement enables a reference-free approach to discover potential quality issues, especially for multilayer vector data, which is the most typical data structure in land mapping and surveying projects. Secondly, the resultant complexity map can provide useful information for QAC in large-scale mapping. For example, efficient sampling plans can be designed on the basis of the complexity map for data inspection and acceptance in terms of spatial stratification and sample arrangement. However, the complexity measurements cannot replace the data inspection process during QAC. Traditional data inspection, such as a manual check or automatic quality check, is still required to determine the actual quality issues with respect to specific spatial features.

Chapter 3 Reference-free method for detecting classification errors in LULC big data

To extend our thought line of reference-free QAC to a single vector layer, a novel method for anomaly detection in LULC big data is demonstrated in this chapter. This method aims to discover classification errors in LULC vector data and provide quantitative estimators for the reliability of the results. To achieve this goal without reference data, a full workflow is designed solely on the basis of vector data and corresponding remote sensing imagery, and a new entropy-based measure is proposed to evaluate the likelihood of a land patch being wrongly classified. This method is expected to benefit the QAC for LULC data with broad geographic coverage, of which the reference data is highly inadequate, as discussed in the Introduction.

3.1 Overview of reference-free methods for LULC data evaluation

Identification and discovery of land classification errors attach great importance to the data revision and uncertainty evaluation during the production of LULC data. The identified results directly determine the accuracy of LULC information and the reliability for further applications (Congalton and Green 2008). Reference data, including existing LULC maps, online geospatial data (e.g. OSM), interpretation by experts and samples collected by fieldwork (Chen and Sun 2014), are widely used as the ground truth in the inspection of LULC data. However, due to the large coverage of LULC big data, full reference data of high quality covering the whole extent of interest are almost impossible (Bruzzone and Marconcini 2009, Olofsson et al. 2013). Sampling inspection is one of the solutions to mitigate that problem by checking selected representative regions (Strahler et al. 2006, Tong et al. 2011, Herold et al. 2014). Nevertheless, although the quantity of required reference data is reduced, the sampling method cannot fundamentally solve the issue of reference dependence, and the errors in those unselected regions are overlooked.

To address the aforementioned problems, researchers have made great efforts in

developing reference-free and reference-reduce methods. On the basis of auxiliary information and internal regularities, a series of reference-free methods have been developed for the uncertainty assessments of VGI (Goodchild and Li 2012, Neis *et al.* 2013, Barron *et al.* 2014, Ali *et al.* 2017). With the aid of remote sensing images, image processing approaches, such as classification and segmentation, are also common in existing reference-free methods (Bruzzone et al. 2004, Baraldi et al. 2005, Bruzzone and Marconcini 2009, Foody 2010, 2012). For example, an ensemble classification method (ECM) was constructed for classification accuracy evaluation by combining observations from multiple classifiers in specific rules (Bruzzone et al. 2004). The outlier detection technique is another option to identify classification errors in current studies. For instance, on the basis of the distribution of the spectral values of land patches, an iterative trimming method (ITM) was designed to detect the statistically outlying patches, which could be taken as the classification errors (Desclée et al. 2006). ITM was later developed with nonparameter estimation and applied to filter reliable samples for global land cover mapping (Radoux and Defourny 2010, Radoux et al. 2014).

The existing methods have alleviated the dependence on traditional reference data in QAC. Nevertheless, the limited quality and quantity of samples for training classifiers will potentially hamper the performance of ECM and other similar methods that highly rely on the training process. The efficiency of such methods also suffers from the mixed composition of land objects and similarity between land classes. In addition, the uncertainty of ITM can be high because ITM are strongly affected by the selection of the probability threshold α, which determines the scope of trimmed data (Radoux and Defourny 2010). Therefore, robust reference-free methods must be further developed.

3.2 Detecting outlying features in land cover data

3.2.1 Underlying assumptions and overall workflow

This study assumes that land data are produced through the general semiautomatic classification, and most parts of the data are correctly mapped. This assumption is reasonable because semiautomatic classification, which compounds the results of visual interpretation and automatic classification, is commonly used in LULC projects and provide acceptable accuracy (Jiang et al. 2012, Lillesand et al. 2014). On this basis, the following extrapolations are obtained:

- Each land class is likely to be dominated by some principal features, such as spectral and textural features. Thus, some outlying features, which can be classification errors or correctly mapped exceptions, may exist. This extrapolation also admits the mixed nature of some land classes (e.g. urban areas).
- Principal features should have a high frequency of occurrence, whereas outlying features are rare. Therefore, principal features are likely to form large clusters in the feature space, whereas outlying features tend to be dispersed from others.

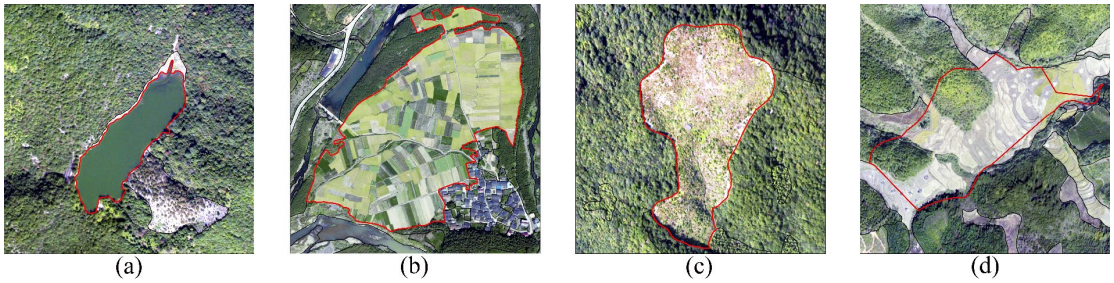


Figure 3.1 (a) Correctly pure ‘water’ object. (b) Correctly mixed ‘paddy’ object. (c) Pure object incorrectly interpreted as ‘building’. (d) Mixed object incorrectly interpreted as ‘grass’.

On the basis of the aforementioned assumptions and extrapolations, mixed and pure land objects, such as the examples shown in Figure 3.1, can be incorrectly

mapped if they contain outlying features that are not normal for the corresponding land class. Therefore, to identify the outlying features, all features within each land class must be initially extracted, and the principal and outlying features can then be divided on the basis of data analysis techniques, such as clustering analysis.

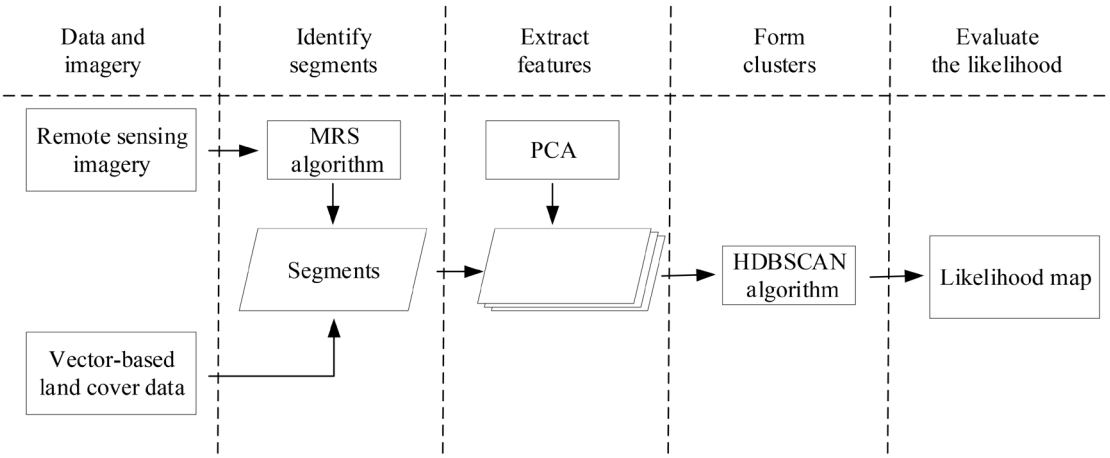


Figure 3.2 Workflow of the proposed approach.

Figure 3.2 illustrates the overall workflow of the proposed method, which comprises four main stages.

- Identify segments. To extract more homogenous segments of each land patch, the land cover data are segmented through a multiple resolution segmentation (MRS) algorithm at a specific scale.
- Extract features. Spectral and textural features are calculated for the extracted segments, and principal component analysis (PCA) is used to obtain a feature space with increased distinguishability.
- Form clusters. The hierarchical density-based spatial clustering (HDBSCAN) is used to cluster the identified features with adaptive clustering parameters. Thus, the features that are dispersed from others can be identified.
- Evaluate the likelihood. On the basis of the clustering results and identified disperse features, a new entropy-based measure is applied to evaluate the likelihood of land patch to be an error. A likelihood map is also generated to show the spatial distribution of potential errors. Details of each stage will be

discussed in the next section.

3.2.2 Determining the segmentation scale for MRS

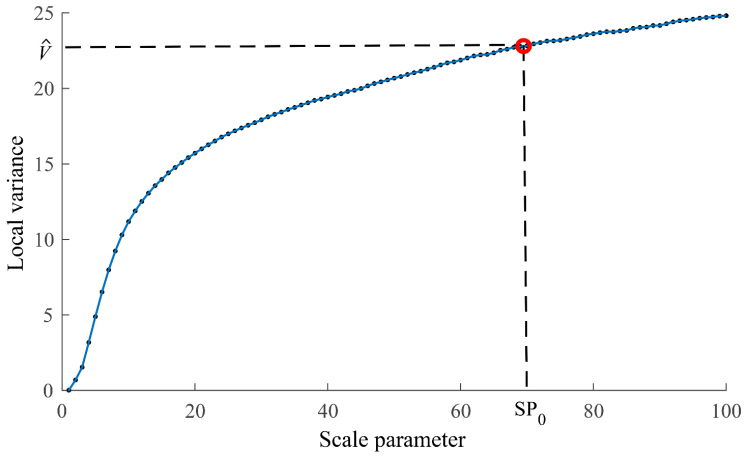


Figure 3.3 Graph of local variance with respect to the scale parameter. The red circle indicates the specific scale SP_0 with an LV of $\hat{V}$.

The segmentation scale, which is determined by the scale parameter (*SP*) in MRS, controls the size and homogeneity of segmentation results (Baatz and Schäpe 2000). To determine an appropriate SP for further evaluation and outlier detection, we use the local variance (*LV*) index, which can be used as an indicator of the scale of each segmentation (Drăguţ et al. 2010), as the indicator to determine the segmentation scale. The segmentation scale is determined when the segmentation result has the same LV estimation to the original map. In this way, we can ensure that the size of ground objects in the segmentation results is basically consistent with the original data. Figure 3.3 shows the graph of *LV* against *SP*. Let $\hat{V}$ indicate the *LV* index for a given land cover data, then the corresponding scale parameter SP_0 can be obtained and further applied for the segmentation process.

The MRS results are further used in identifying the original land patches to maintain the land information of original land cover data in the further analysis. Tiny segments are removed because they can introduce insignificant features. As shown in Figure 3.4, the segments smaller than the MMU are merged to the neighbour that has the same class label and shares the longest boundary.

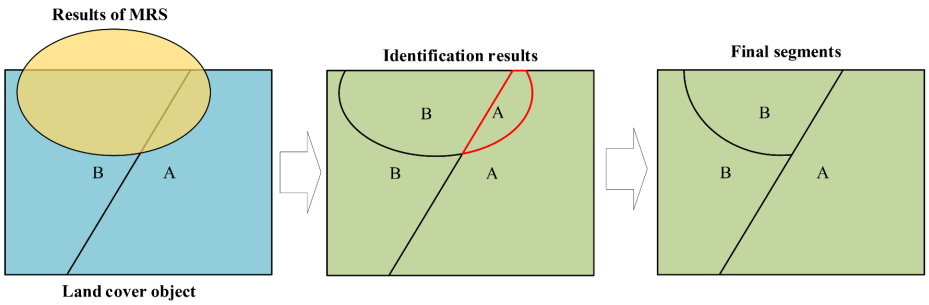


Figure 3.4 Procedure for obtaining the final segments. The area within the red line is smaller than the MMU.

3.2.3 Feature extraction and clustering

Spectral and textural features are extracted for each segment generated in the last session. The spectral feature includes the mean and variance of the grey values within a segment. Textural features are computed based on a grey-level co-occurrence matrix in a 5×5 window. Given that some texture features are strongly correlated (Wang and Zhang 2014), only four textural features are selected, including the variance, contrast, homogeneity and second moment. Therefore, for an RGB remote-sensing image that has three visible bands, a 30-dimension feature space is constructed.

- Mean grey value of RGB bands (three dimensions)
- Variance of the grey value of RGB bands (three dimensions)
- Mean texture of RGB bands (12 dimensions)
- Variance of the texture of RGB bands (12 dimensions)

The significance of each feature may differ for different land classes. For example, the spectral feature attaches great importance in distinguishing an area with vegetation, whereas the textural feature might be more useful for identifying an urban area. Therefore, PCA is conducted before clustering to select the most significant features (Demšar et al. 2013). In this study, the first three principal components of PCA are taken for further clustering.

To identify the outliers amongst the extracted features, the HDBSCAN algorithm is used to cluster the points in the projected PCA space. HDBSCAN is an exploratory clustering algorithm that can obtain stable clusters without prior knowledge about the

data. HDBSCAN requires fewer intuitive clustering parameters and is claimed to have better performance than other clustering methods (e.g. DBSCAN, K-means and MeanShift; McInnes and Healy 2017).

The clustering parameter m_{clsize} denotes the minimum cluster size and strongly affect the results of HDBSCAN. To make the parameter m_{clsize} adaptive to different situations, a loose criterion is made for the selection of m_{clsize}. Multiple values of m_{clsize} are tested, and the value that produces the highest clustered degree is selected as the optimal value. Under that loose criterion, the nonclustered data can be safely taken as outliers. The clustered degree R_{cls} of a feature set with $m_{\mathrm{clsize}} = m$ is defined as follows:

$$R_{\mathrm{cls}}(m) = \frac{F_{\mathrm{cls}}(m)}{F_{\mathrm{all}}} \times 100\% , \tag{3.1}$$

where $F_{\mathrm{cls}}(m)$ is the number of features assessed as clustered, and F_{all} indicates the total number of features.

On the basis of the theoretical basis of HDBSCAN algorithm and our preliminary experiments, R_{cls} tends to be excessively large when m_{clsize} is extremely small (e.g. $m_{\mathrm{clsize}} = 2$; McInnes and Healy 2017). To address this problem, an upper limit (i.e. 95%) is set for R_{cls}, and the m_{clsize} that produces a R_{cls} higher than this limit will not be considered in this study. After clustering, all segments will be given a cluster label on the basis of the clustering results of their own features, and the segment that corresponds to outlying features are considered a single cluster and assigned to a unique cluster label.

3.2.4 Likelihood measure: Design and rationality

The likelihood of a land patch to be an error is measured on the basis of the diversity and abnormality of the features within the patch. Diversity represents the degree of mixing of a patch, whereas abnormality refers to the proportion of potential incorrectly mapped part of a land patch.

Then, the diversity of a land patch can be measured in terms of ENT (Shannon and Weaver 1949). Therefore, the diversity of a land patch X can be expressed as follows:

$$H(X) = -\sum_{i=1}^{n} p_i \ln(p_i) , \tag{3.2}$$

where p_i denotes the proportion of the area covered by the segments with cluster label i, and n indicates the number of cluster labels within patch X.

Let S_i represent the set of segments with cluster label i, and A_{S_i} be the sum area of S_i. On the basis of our assumption in Section 4.2.1, a high A_{S_i} indicates the globally broad coverage of the items in S_i and implies a high possibility for these items to be correctly mapped. Therefore, these items will have a low contribution to diversity estimation. As a result, Equation (3.2) is further modified as

$$H(X) = -\sum_{i=1}^{n} (1 - \frac{A_{S_i}}{A}) p_i \ln(p_i) , \tag{3.3}$$

where A is the total area covered by the land class of patch X. Mathematically, $H(X)$ reaches its maximum when $A_{S_i} = 0$ and p_i equals $1/n$, that is,

$$H(X)_{\max} = \ln(n) . \tag{3.4}$$

Therefore, to support the comparison between patches, the normalised diversity measurement can be expressed as follows:

$$H(X)_{\text{norm}} = \frac{H(X)}{H(X)_{\max}} . \tag{3.5}$$

Furthermore, the abnormality of a patch X is represented by the proportion of area covered by outlying segments. Similar to diversity measurement, the coverage of outliers indicates their significance. Thus, the abnormality index can be expressed as follows:

$$B(X) = (1 - \frac{A_{\text{outlier}}(X)}{A}) \frac{A_{\text{outlier}}(X)}{A_X}, \tag{3.6}$$

where A_X and $A_{\text{outlier}}(X)$ represent the area of patch X and the outlying segments within X, respectively.

By combining Equations (3.5) and (3.6), the likelihood for a patch to be an error can be expressed as follows:

$$L(X) = H(X)_{norm} + B(X). \tag{3.7}$$

3.3 Experiments and analysis

3.3.1 Data description and experimental setup

Land cover datasets with different mapping scales are selected for the experiments. The first dataset (MMU is 400 m^2) is collected in the project of National Geographic State Monitoring (NGSM) in China. The selected dataset contains 6,016 land patches into 38 land classes and covers a region of $12{,}000 \times 8{,}200$ m^2 in Fujian, China. The image corresponding to this dataset was captured by unnamed aerial vehicles in July 2012 with a spatial resolution of 2 m. The other dataset (MMU is 250,000 m^2) is selected from the project of Corine Land Cover (CLC) 2012. The dataset consists of 1,444 land patches categorised into 23 classes with an area of $60{,}000 \times 60{,}000$ m^2 in Miskolc, Hungary. The Rapideye image acquired in August 2018 with a spatial resolution of 5 m is used for this dataset (Planet Team 2017).

To validate the results of the detected errors, the ground truth is generated through a visual inspection. The results of the inspection indicate that the NGSM dataset has 214 errors, of which the inside features are significantly different from those of other patches belonging to the same land class. For the CLC dataset, 54 errors are identified.

On the basis of the process described in Section 4.2.2, the *SP* values for NGSM and CLC datasets are computed to be 82 and 220, respectively. The clustering parameter m_{clsize} is iteratively tested from 2 to 30, whereas the upper limit for the

rate R_{cls} is set to 95%. The ground truth, together with the raw vector data and remote sensing images, are shown in Figure 3.5(a) and Figure 3.6(a) for those two datasets. The corresponding likelihood measurements are shown in Figure 3.5(b) and Figure 3.6(b).

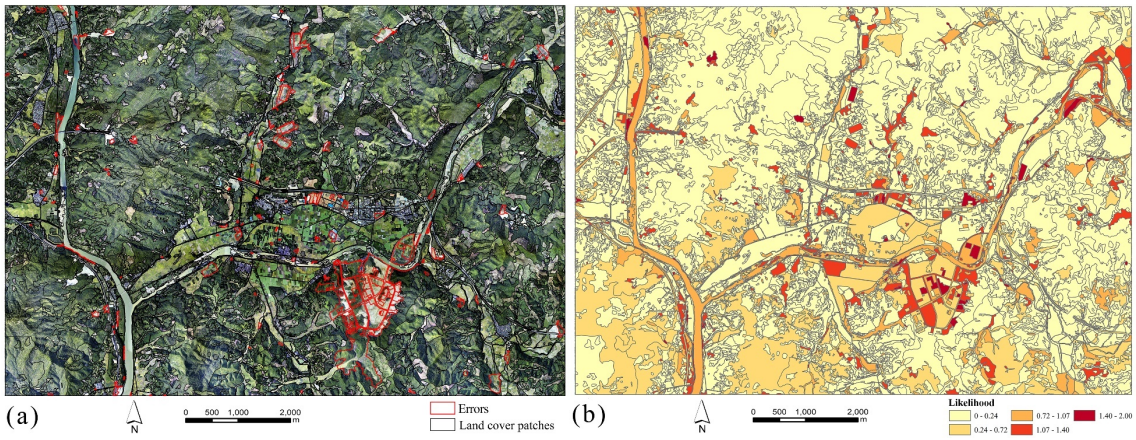


Figure 3.5 NGSM dataset. (a) Land cover, errors and image. (b) Likelihood map.

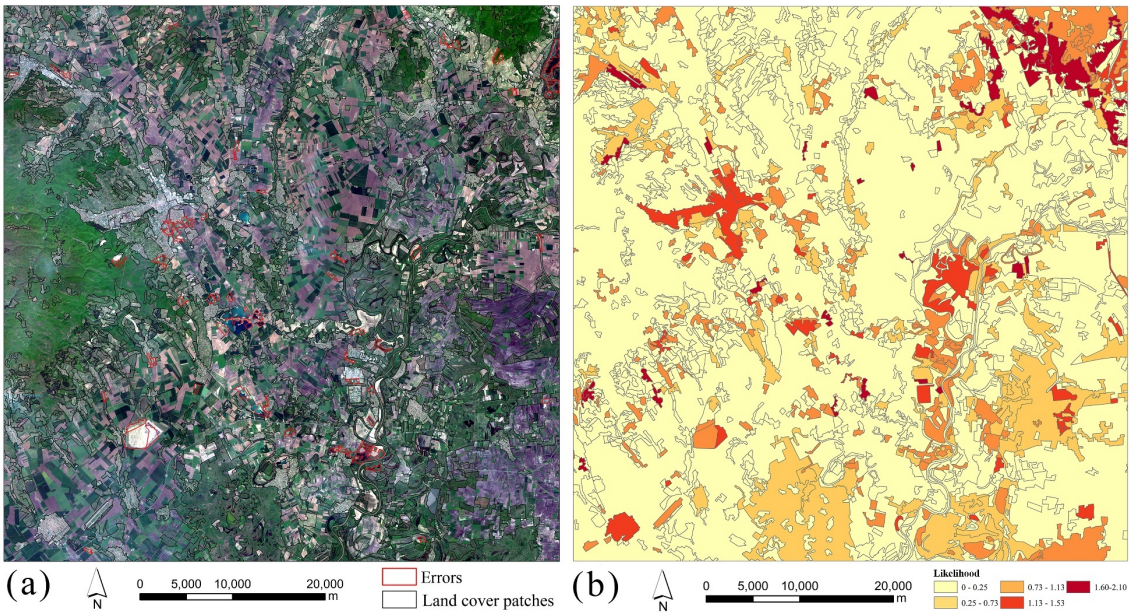


Figure 3.6 CLC dataset. (a) Land cover, errors and image. (b) Likelihood map.

3.3.2 Comparison with state-of-the-art methods

As discussed in Section 4.2.4, patches with a high likelihood measurement are expected to consist of outlying features or possess a high degree of mixing. These patches will be taken as 'suspect' (i.e. potential errors), whereas the other patches (i.e. likelihood equals to 0) will be considered as 'trusted' because they are pure and

consistent within their classes in terms of spectral and textural features.

The performance of the proposed method is compared with state-of-the-art methods. Representative methods based on classification analysis and outlier detection technique, including ECM and ITM, are selected for comparison. For ECM, three classifiers, namely, K-nearest neighbour with k equals 1, decision tree and Bayes classifiers, are used. The first half area patches (ascending order by area) of each class are selected for training the classifiers. The patches that have a different classification from the result of ECM are assessed as 'suspect'. For ITM, the segmentation and image features used in the proposed method are inherited for a fair comparison. The land patches that contain outlying segments are assessed as 'suspect'. Multiple probability thresholds α between 20% and 40% are tested.

Table 3.1 Results of different methods on NGSM dataset.

Model	N_S	N_{TPE}	**TPR**	**PPV**
Proposed method	1232	153	71.50%	12.42%
ECM	2511	139	64.95%	5.54%
ITM, α=0.2	927	81	37.85%	8.74%
ITM, α=0.3	1660	131	61.21%	7.89%
ITM, α=0.4	2294	151	70.56%	6.58%

Table 3.2 Results of different methods on CLC dataset.

Model	N_S	N_{TPE}	**TPR**	**PPV**
Proposed method	424	44	81.48%	10.38%
ECM	397	26	48.15%	6.55%
ITM, α=0.2	140	18	33.33%	12.86%
ITM, α=0.3	287	26	48.15%	9.06%
ITM, α=0.4	434	36	66.67%	8.29%

Two metrics, namely, true positive rate (TPR) and precision (PPV), are used to evaluate the accuracy of each method. $\mathrm{TPR} = N_{\mathrm{TPE}}/N_{\mathrm{TE}}$, where N_{TPE} refers to the number of true errors within the 'suspect' group, and N_{TE} is the total number of true errors. $\mathrm{PPV} = N_{\mathrm{TPE}}/N_{\mathrm{S}}$, in which N_{S} is the number of 'suspect' patches. High values of the two indexes represent excellent performance in error discovery.

As shown in Table 3.1 and Table 3.2, the proposed method demonstrates its good

performance by scoring the highest TPR (i.e. 71.50% and 81.48% for NGSM and CLC datasets, respectively). The ITM with $\alpha = 0.4$ has a similar TPR of 70.56% for NGSM dataset, whereas the corresponding PPV (6.58%) is considerably lower than that of the proposed method (12.42%), indicating the higher false detection rate of ITM in this case. Although the PPV of ITM with $\alpha = 0.2$ scores the highest value of 12.86% amongst other methods in the case of CLC dataset, the corresponding TPR is the lowest of only 33.33%, which presents the poor performance of detecting true errors. PPV and TPR vary in relatively wide ranges. Thus, the ITM shows significant sensitivity to the selection of parameter α, which can result in significant uncertainty in practical application. For the ECM method, the PPV for the two datasets is only 5.54% and 6.55%, which are significantly lower compared with other methods. The poor performance of ECM can be explained because certain land categories are conceptually close and show high similarity in terms of spectral and textual features (e.g. 'broad-leaf forest' and 'coniferous forest'), which increases the uncertainty during training classifiers and the following classification.

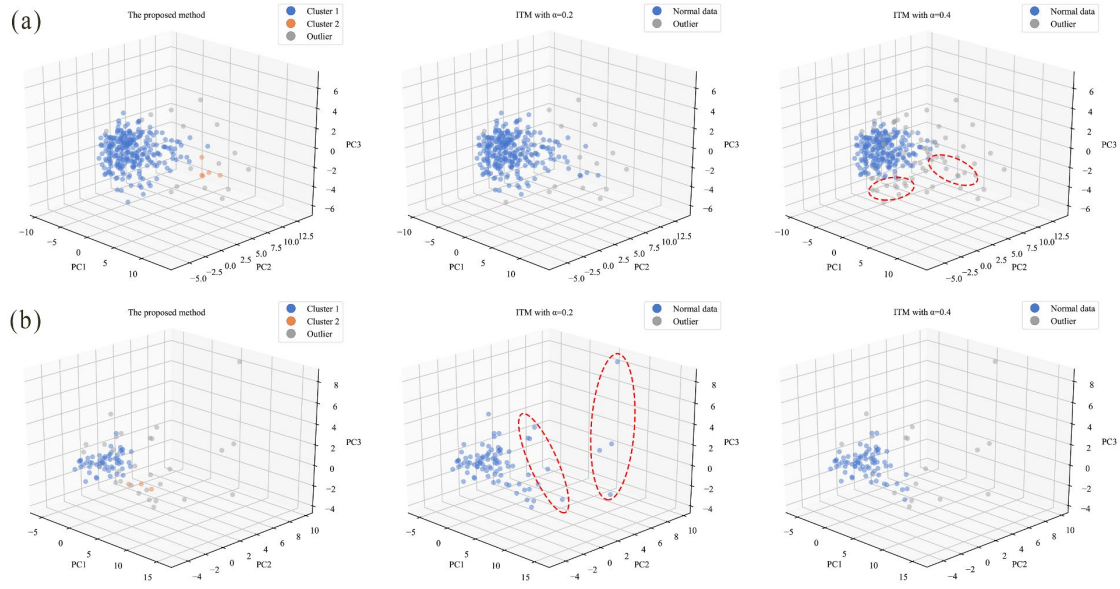


Figure 3.7 Outlier detection results of two land classes in CLC dataset. (a) Transitional woodland shrub. (b) Water bodies. PC1, PC2 and PC3 denote the first three PCA components.

To further investigate the advantages in outlier detection of the proposed method,

the clustering results of two land classes in CLC dataset, namely, 'transitional woodland shrub' and 'water bodies', are illustrated and compared with the those of ITM in Figure 3.7(a) and Figure 3.7(b), respectively. For 'transitional woodland shrub' in Figure 3.7(a), two clusters are identified through the proposed method, and the rest ones are taken as outliers. Similar detection results are obtained by ITM with $\alpha = 0.2$, whereas with $\alpha = 0.4$, certain points are wrongly detected as outliers, as marked by the red circle. For 'water bodies' in Figure 3.7(b), certain significant outliers (marked by the red circle) are missed in ITM with $\alpha = 0.2$, whereas similar and better detection results are obtained by the proposed method and ITM with $\alpha = 0.4$. This comparison shows that the proposed method outperforms ITM in terms of robust performance, and it is less reliant on initial parameters, which is attributed to the adaptive parameter setting during the clustering process.

3.3.3 Effectiveness of likelihood measurement

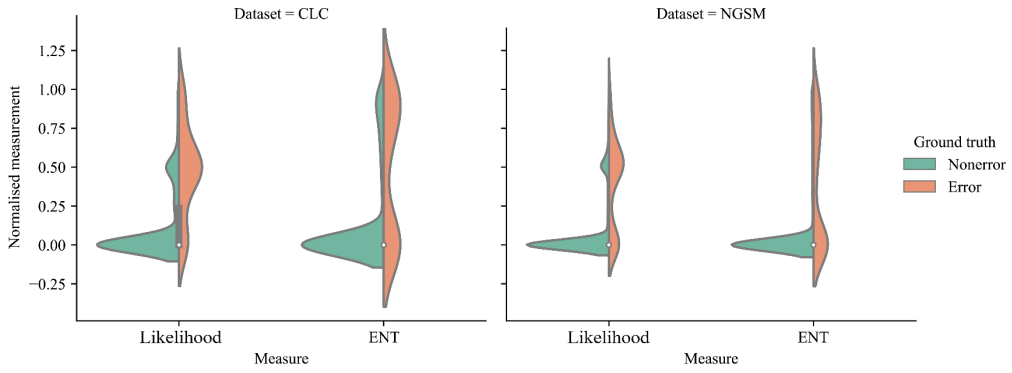


Figure 3.8 Violin plots of normalised likelihood and ENT for different patch groups.

To evaluate the effectiveness of the quantitative description of the likelihood measurement, violin plots of the normalised likelihood and ENT (normalised through the division with the maximum) for the two datasets are illustrated in Figure 3.8. Visually, the real errors are likely to have large likelihood values, whereas the likelihood for non-errors tends to be small. ENT shows similar effectivity in distinguishing non-errors, whereas a large portion of real errors is assigned with small ENT values. This finding can be explained by the fact that the ENT can only reflect

the degree of diversity while is theoretically incapable of capturing erroneous patches if they are pure. In this sense, the proposed likelihood measure outperforms the ENT by providing better quantitative descriptions for distinguishing errors and non-errors.

In the previous comparison test, 0 is used as a threshold for the likelihood in assessing a patch as 'suspect' or 'trusted'. However, the violin plots indicate that increasing this threshold may help reduce the portion of non-errors in the 'suspect' group. To validate this hypothesis, the curves of PPV and TPR with respect to different thresholds are plotted in Figure 3.9. For the two datasets, as the threshold increases, the PPV keeps growing, whereas the TPR undergoes a significant decline, especially after the threshold reaches 0.4. Therefore, 0.4 can be a suitable threshold for the normalised likelihood when applied to these datasets. The effectivity of the likelihood measurement will be enhanced by such threshold, given that many non-errors can be excluded from the 'suspect' group, whereas most true errors are retained.

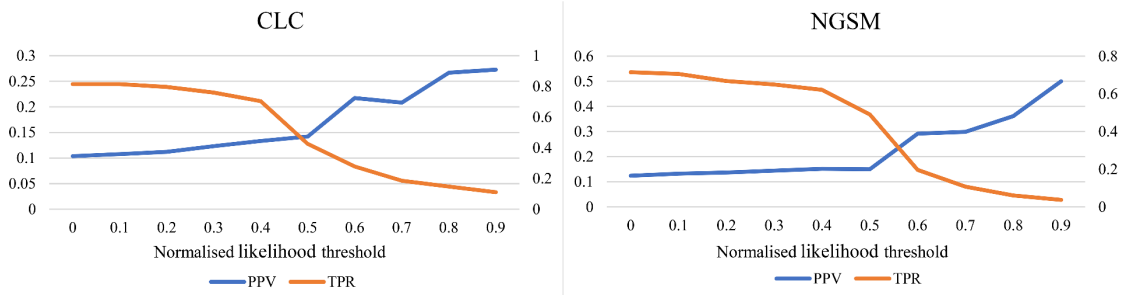


Figure 3.9 Curves of PPV and TRP with respect to different thresholds.

3.4 Summary

This chapter demonstrated a reference-free classification error detection method for LULC data. In this method, an adaptive segmentation analysis process was initially conducted to divide the given land patches into homogenous segments. Then, spectral and textural features were extracted for those segments, and an adaptive clustering strategy was designed to discover outlying segments. Finally, a new entropy-based measure was developed to measure the likelihood of a land patch being a classification error. This method is expected to benefit the QAC in LULC big data because it can automatically discover potential classification errors without any reference data and

provide a quantitative evaluation of these errors. The evaluation results will be reported to users for further inspection and possible amendments to improve the final classification result. However, even though the proposed method seems to perform better than classic methods, there is still much space for improvement concerning its low PPV values. In that sense, the ensemble learning technique that combines the power of multiple outlier detectors would be a promising direction in our future work. Also, the spectral and textural features used in this study might not be effective for all types of land objects, and the uncertainty raised by manual selection of features would also affect the performance of outlier detection. As the concept of deep feature extracted by artificial intelligence has been widely adopted in image classification and segmentation (Romero *et al.* 2015, Chen *et al.* 2016), it would be one of the potential solutions to enhance the feature extraction and produce more distinguishable features in terms of different land objects.

Chapter 4 Reference-free method for investigating scale uncertainty in LULC big data

A data production specification is always predefined for the collection of LULC data, which regulates some technical details, such as the definition of land classes and MMUs. The production specification indicates the degree of abstraction of the land surface and will inevitably introduce some uncertainties. In this chapter, the effect from MMU, which determines the mapping scale in LULC project, is studied, and a reference-free method is proposed to evaluate the uncertainty caused by MMU.

4.1 Overview of scale uncertainty in LULC

Land classification system (LCS) is one of the most critical components that provide standards for mapping and classification of LULC data (Di Gregorio 2005). An LCS always has multiple levels of land classes, and these levels are designed in a nested structure in terms of the degree of detail of the corresponding spatial object (Wyatt et al. 1997). For example, the class of 'water' may include subclasses 'river' and 'lake'. Moreover, an LCS defines the classification criteria and mapping rules for each land class, which has a direct influence on the process of image interpretation and data collection in practice (European Environment Agency 2007). Theoretically, LCSs are only an abstracted version of the real world, and uncertainty will be inevitably introduced in the resultant data, as well as the analyses conducted on the data (Verheye 2009).

Generally, LCS has two primary uncertainty sources. The first source is the definition of land classes. However, the defined classes cannot be absolutely exhaustive and conceptually exclusive (Anderson 1976, Jansen et al. 2000); thus, semantic uncertainty exists, especially when different LCSs are applied to the same datasets (Feng et al. 2004, Xu 2016). The other source is the mapping scale, which is used to enhance the efficiency in representing the real world (Di Gregorio et al. 2012). The mapping scale determines the degree of land details that will be considered in data

production and inevitably leads to information loss.

The scale is an essential issue in the study of GISci and remote sensing techniques (Cao et al. 1997, Goodchild et al. 1997, Benz et al. 2004). Myint et al. (2011) reported that the classification accuracy gradually declines with the increase of the scale level. Therefore, understanding and even qualifying the effect of the mapping scale before the production of LULC data is necessary to control the uncertainty in the process of classification (Wu et al. 2006). For LULC data production, the mapping scale is always specified by the MMU, which determines the smallest land patch size allowed being mapped (Lillesand et al. 2014). In practice, the land patches smaller than MMU will be removed and merged into their neighbours. An extremely large MMU will lead to oversimplification issue because many land details are missed (Saura 2002).

In practice, MMUs for different classes are commonly set to empirical values (EEA 1995, Rahman et al. 2001, Fry et al. 2008). However, on the basis of previous studies, the selection of MMUs may considerably affect the spatial configuration and statistical information of LULC data (Saura 2002, Pascual-Hortal et al. 2007, Rutchey et al. 2009, Kelly et al. 2011). Patches smaller than MMU will be omitted in the LULC production and ultimately generate some unexpected classification errors. As a result, an evaluation is needed to measure the uncertainty and support the modifications of MMUs. A simple evaluation directly compares the current data with the reference data, which are produced without the restriction from MMUs. However, the mapping process is irreversible and labour-intensive; thus, obtaining the reference data is difficult. Hence, the feasibility of this approach is always in question.

4.2 Modelling omission errors

4.2.1 Assumption on the skewed distribution of patch size

The omission errors caused by MMUs are derived from the elimination of small land patches. These small patches will be merged into their neighbours, and the attribute values will also be modified into the neighbours. Some scholars have indicated that

the number of small patches tends to be considerably higher than the number of large ones, whereas these small patches only account for approximately 5% of the total area (Wang et al. 1995). Therefore, we assume that the frequency of patch size should obey a positively skewed distribution, in which more patches tend to have a small size.

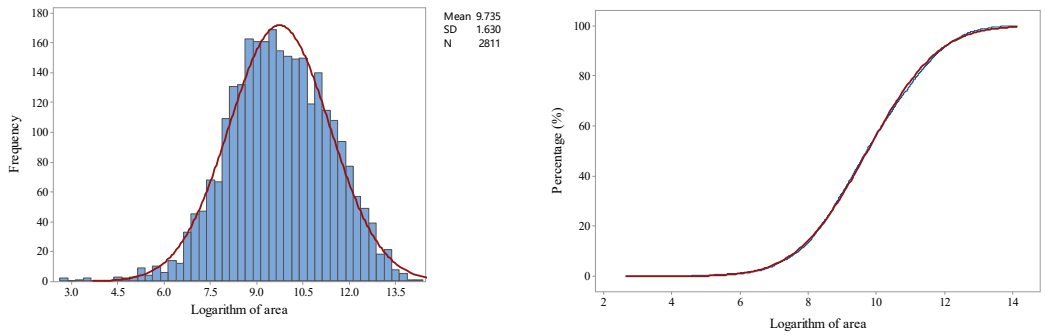
Table 4.1 Skewness and kurtosis for land use samples in the selected region.

Class name	**Raw data**		**Log-transformed data**	
	Skewness	Kurtosis	Skewness	Kurtosis
Forest	23.02	950.17	0.80	0.73
Park	58.05	3,691.89	−0.28	0.26
Residential	5.66	54.76	0.16	−0.94
Industrial	18.03	526.47	−0.47	0.51
Farm	149.34	26,474.95	0.34	1.32
Cemetery	11.06	171.80	0.07	0.99
Allotment	16.26	509.61	0.39	0.23
Meadow	82.66	9,048.35	−0.07	0.35
Commercial	4.67	31.21	−0.18	0.09
Nature reserve	4.99	36.81	−0.48	0.15
Recreation ground	10.17	147.11	−0.86	0.72
Retail	7.6	84.04	−0.35	2.29
Military	11.09	127.44	−0.5	0.12
Quarry	5.02	40.26	−0.67	0.02
Orchard	85.07	10,194.69	0.59	0.7
Vineyard	12.12	240.2	0.59	0.27
Scrub	24.97	1,084.64	−0.22	0.78
Grass	96.59	12,398.02	−0.10	-0.04
Heath	12.26	217.94	0.54	-0.14

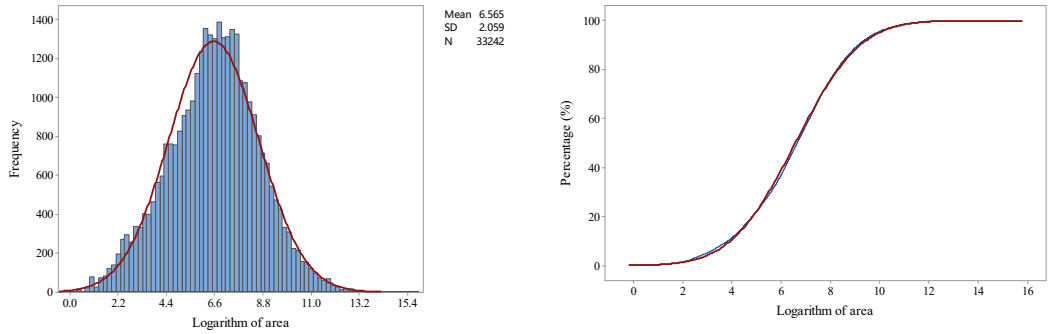
Many pieces of evidence can be found to support the assumption in previous research (Dunajski et al. 2008, Sharma et al. 2011, Filstrup et al. 2014). However, to make the assumption convincing, the land use data from the OSM project (http://download.geofabrik.de/) are selected for additional validation. The dataset of Baden-Württemberg, Germany, which has been proven to be of better data quality than the other regions, is selected (Neis et al. 2014). The selected dataset consists of 19 classes and contains 407,765 land patches in total. Notably, no specific limitation exists on patch size during the land use collection in the OSM project. Therefore, the

patch size can be taken as a relatively full representation of the real world.

In this study, a visual validation is conducted on OSM data by applying a logarithmic transformation, which is commonly used in normalising skewed distributions (Zhang et al. 1998), to the area of land patches. As shown in Table 4.1, the logarithmic transformation brings a significant improvement in the statistics of skewness and kurtosis. For the data after transformation, the kurtosis values become close to 0, and most values of skewness are between −0.5 and 0.5, which indicates that the dataset can be normally distributed (Saunders 1981). Furthermore, Figure 4.1 presents the histogram and empirical cumulative distribution function (CDF) of the logarithmically transformed data for the land use classes ‘commercial’ and ‘grass’, indicating the potential positively skewed distribution of the raw data (O Ztuna et al. 2006).



(a) Histogram and empirical CDF of ‘Commercial’ class



(b) Histogram and empirical CDF of ‘Grass’ class

Figure 4.1 Histogram and empirical CDF of the logarithmically transformed data of the ‘Commercial’ and ‘Grass’ class in Baden-Württemberg, Germany. SD denotes the standard deviation of the fitted normal distribution.

4.2.2 Approach for the prediction of omission area

An example based on the assumption on skewed distribution is shown in Figure 4.2. The left tail is truncated due to the limit of MMU, which discards those patches belonging to the shadowed part and thereby introduces omission errors. To recover and calculate the amount of these omission errors, the distribution of the original data should be initially identified.

To address the above issue, a set of transformations are designed and applied to the raw data. Then, the best transformation is selected in terms of the goodness of the transformed result in approximating a normal distribution. Finally, the original distribution can be recovered through inverse transformation.

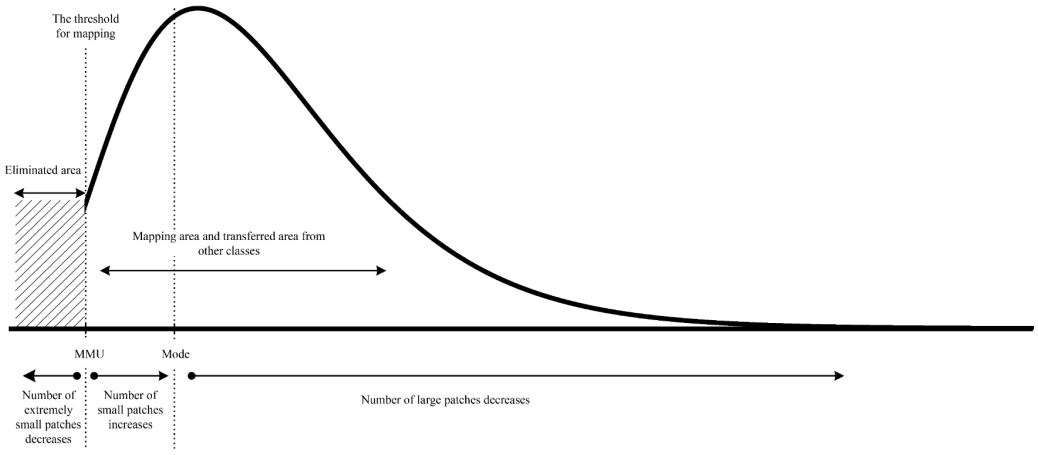


Figure 4.2 Effect of MMU on patch size distribution.

The transformations used in our method are designed based on Box-Cox transformation, which is effective in correcting the nonnormality of data (Box et al. 1964). The Box-Cox transformation for variable y can be expressed as

$$W=\begin{cases} y^{\lambda} & \lambda \in [-1,0)\cup(0,1] \\ \ln(y) & \lambda = 0 \end{cases}. \tag{4.1}$$

However, in case of some complex skewed distributions, Equation (4.1) is redesigned as a power function that has high adaptivity, and the power function can be written as

$$W(g(y),\lambda) = g(y)^{\lambda} \qquad \lambda \in [-1,0) \bigcup (0,1]. \tag{4.2}$$

Theoretically, the term $g(y)$ in Equation (4.2) can be defined as any function that can be used to correct the positively skewed distribution, such as $\mathrm{g}(y) = y^{0.5}$, $\mathrm{g}(y) = \ln(y)$ or $\mathrm{g}(y) = y^{-1}$. However, the base function $g(y)$ should not be extremely complicated to simplify the computation and make the transformations explainable. Notably, only a finite number of the exponent λ will be tested in practice. Otherwise, the transformations will be infinite. Therefore, in this study, the base function $g(y)$ is defined as three basic functions, that is, $\mathrm{g}(y) = y$, $\mathrm{g}(y) = ln(y)$ and $\mathrm{g(y)} = \mathrm{y}^{-1}$. The exponent λ is set to $\{0.1, 0.2, \ldots, 1\}$. As a result, 30 default transformations into three sets are designed as follows for further analysis:

$$W = \begin{cases} y^{\lambda} \\ \ln(y)^{\lambda} \\ y^{-\lambda} \end{cases} \qquad \lambda = \{0.1, 0.2, ..., 1\}. \tag{4.3}$$

The full distribution cannot be directly modelled Due to the blank at the left tail of the transformed data, (Law 1991). In this study, the curve fitting technique is used to approximate the distribution of transformed data (Motulsky et al. 2004). The transformed data are initially binned into equally spaced containers. The number of containers N_{bins} is limited to the interval from 30 to 1,000. Within that interval, N_{bins} is selected as large as possible, provided that more than 80% containers are with data. Subsequently, a normal distribution function (Equation (4.4)) is used to fit the binned data, where μ, σ and A are the parameters to be identified. In this study, the goodness of fitting results is measured as the sum of squares for error (SSE), and the best transformation function is denoted as $W(g_0(y), \lambda_0)$.

$$f = A \times \mathrm{e}^{-\frac{(y-u)^2}{2\sigma^2}} \tag{4.4}$$

Finally, the omission area E_{o} can be recovered with the parameters of the best transformation, which can be expressed as follows:

$$E_{\mathrm{o}} = F_{\mathrm{sum}} \int_{0}^{MMU} \left[y \times \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{[W(g_0(y),\lambda_0)-u]^2}{2\sigma^2}} \times W_y(g_0(y),\lambda_0) \right] \mathrm{d}(y), \tag{4.5}$$

where F_{sum} denotes the predicted total frequency. Particularly, F_{sum} is computed as a function with respect to the observed frequency F_{o}, as shown as follows:

$$F_{\mathrm{sum}} = \frac{F_{\mathrm{o}}}{\int_{\mathrm{MMU}}^{+\infty} \left[\frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{[W(g_0(y),\lambda_0)-u]^2}{2\sigma^2}} \times W_y(g_0(y),\lambda_0) \right] \mathrm{d}(y)}. \tag{4.6}$$

4.2.3 Validation test on simulated and real-life data

The validation of the prediction method is conducted in the following processes:

(1) Randomly generate a set of normal variable $X \sim N(u, \sigma^2)$ $(X > 0)$ with a size of n.

(2) Obtain a skewed distributed variable $c(X)$ by applying a specific conversion function $c(x)$, such as $c(x) = \mathrm{e}^x$, to variable X.

(3) Remove the smallest part of the skewed data at a specific percentage P and count the lost area $\hat{E} = \sum c(X)$ $(X < X_P)$, where X_P indicates the variable value at P.

(4) Calculate the accuracy rate $R = 1 - \left|\hat{E} - E_{\mathrm{o}}\right| / \hat{E}$ as the metric for the goodness of prediction. The rate R represents the prediction accuracy and will become more significant as the predicted omission area E_{o} becomes closer to the true value $\hat{E}$.

A set of simulation tests is made with different inputs to validate the proposed processes. Given the random error in randomly generated normal variables, we repeat each simulation 100 times to reduce random uncertainty. Figure 4.3 shows the outputs of the simulation tests. The performance of the proposed prediction approach are compared with different datasets, as shown in Figure 4.3(a). In most cases, the accuracy rate R is higher than 0.7 with different the initial conversion functions and values of discarded percentages P. As shown in Figure 4.3(b), the boxes for $n = 500$

generally span a large interval with a lower median value than those for $n = 5{,}000$, which indicates that enhanced performance is based on large data size. Figure 4.3(c) and Figure 4.3(d) show the effect of data size. As the data size becomes smaller (e.g. $n = 300$), the accuracy rate ranges within a large interval (Figure 4.3(c)). The average of accuracy rate R obtains a significant improvement when the data size increases to 1,000, and after 1,000, the rate becomes relatively stable (Figure 4.3(d)).

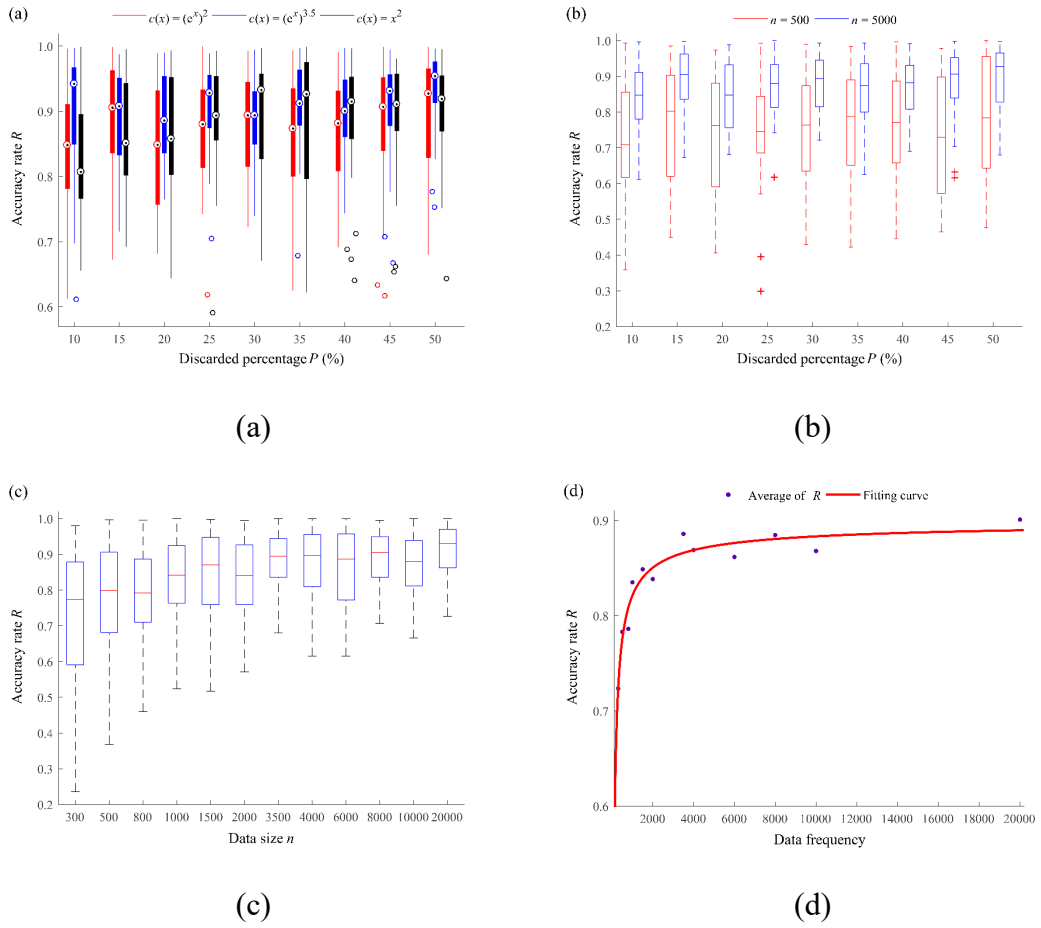


Figure 4.3 Results of simulation tests. The raw data are generated with parameters μ=3, $\sigma = 1$ and $X \in (0,6)$. (a) Boxplots of the accuracy rates with different initial conversions. The data have a size n of 5,000. (b) Sensitivity analysis on data size. $c(x) = (e^x)^2$. (c) Effect of data size on accuracy rate. The discarded percentage P is 30%, and $c(x) = (e^x)^2$. (d) Average accuracy rate with different data sizes. A power fitting curve $a * x^b + c$ is applied.

The real-life OSM data used in the last section are selected for further validation. Land patches smaller than the given MMU, which is set to 3,000 m^2, are removed, and the remaining data are used for the recovery. As shown in Table 4.2, the accuracy rate

R is higher than 0.7 for classes (i.e. indexes 6–19) with large data size (i.e. frequency above 1,000), whereas the accuracy significantly decreases for those with small data size (i.e. indexes 1–5).

Table 4.2 Prediction accuracy of OSM data. The real omission area is the sum of discarded patches, and the frequency is the number of rest patches.

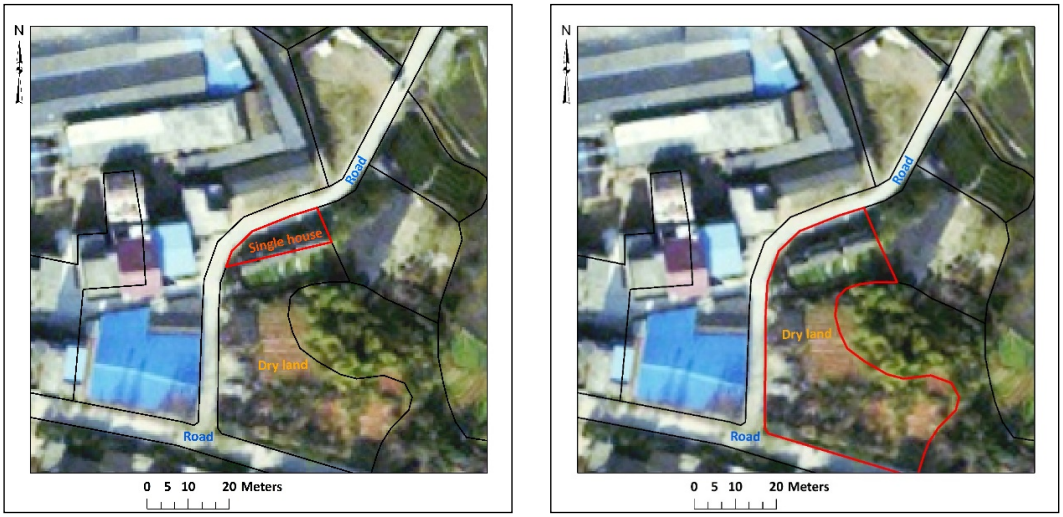
Data size	Class index	Class name	Frequency	Predicted omission (m^2)	Real omission (m^2)	R
Small	1	Military	118	14,732	15,191	0.970
	2	Retail	457	116,628	77,108	0.487
	3	Nature reserve	588	1,755	36,725	0.048
	4	Quarry	737	61,879	167,818	0.369
	5	Recreation ground	813	370,820	468,265	0.792
Large	6	Commercial	1,352	1,391,326	2,226,869	0.625
	7	Cemetery	2,142	2,090,615	2,330,388	0.897
	8	Heath	2,230	2,224,496	2,212,994	0.995
	9	Park	2,432	753,250	590,535	0.724
	10	Industrial	4,567	728,987	899,673	0.810
	11	Allotments	5,815	7,788,601	7,800,704	0.998
	12	Vineyard	7,124	7,999,473	7,961,102	0.995
	13	Residential	7,921	9,813,440	18,114,947	0.542
	14	Orchard	13,120	5,887,513	5,133,799	0.853
	15	Grass	15,005	20,440,658	19,547,255	0.954
	16	Forest	15,569	24,657,985	39,807,563	0.619
	17	Scrub	29,696	28,477,379	22,658,485	0.743
	18	Meadow	51,102	38,909,807	42,208,925	0.922
	19	Farm	80,835	12,253,330	17,888,881	0.685

The validation results of simulated and real-life data show that the performance of the proposed method is well and robust for large data size; whereas it may not be efficient for small data size. Therefore, to control the uncertainty brought by the small data size, the minor LULC classes, of which the number of patches is less than 1,000, will not be involved in the prediction process in this study. Notably, given the small data size of these minor LULC classes, the omission errors may be negligible; thus, the overall accuracy will not be significantly affected. For this reason, even these minor classes also have omission errors, and these errors will be reasonably ignored in the evaluation.

4.3 Modelling commission errors

4.3.1 Practical generalisation rules for merging small patches

The generalisation rule determines how data will be abstracted in LULC production (Cheng et al. 2006). In practice, tiny patches that are smaller than a specific MMU will be merged into its neighbours to ensure that land surface will be fully covered by LULC patches. As a result, commission errors will be introduced by the merging operation. To measure the commission errors, the possibility for one specific class conveying into another must be estimated. In this study, we select the most widely used generalisation rule in commercial GIS software as an example to demonstrate how commission errors can be modelled. The selected generalisation rule regulates that the tiny LULC patches will be merged into one of its neighbours that has the /longest shared boundary (Wang et al. 1996). Notably, although the generalisation rule can be different in practical projects, the idea of this proposed method can be extended because the calculation process is easily modifiable.



(a) Before merging (b) After merging

Figure 4.4 Instance of merging operation. The 'Single house' patch will be merged into the 'Dry land' neighbour which has the second-longest shared boundaries because of its small area and the relatively rigid boundary of 'Road'.

Notably, some classes, such as 'road', always have long-shared boundaries with the adjacent patches due to their network structure across the land surface. However,

these classes that have network structures may be unlikely selected as the target in the merging operation because of their relatively rigid boundary and unique shape. For this reason, these classes will be excluded from the merging operation. For example, as shown in Figure 4.4, the patch 'road' in (a), which needs to be removed due to its smaller area than MMU, shares the longest boundary with patch 'single house' and the second-longest boundary with patch 'dry land'. However, the patch 'single house' will be merged into 'dry land' because 'road' class is excluded.

4.3.2 Computing commission errors based on the conversion between land classes

On the basis of the generalisation rule, the length of shared boundaries between land classed can be used to estimate the conversion possibility between them. On this basis, the conversion possibility p_{ij} from class j to class i can be expressed as

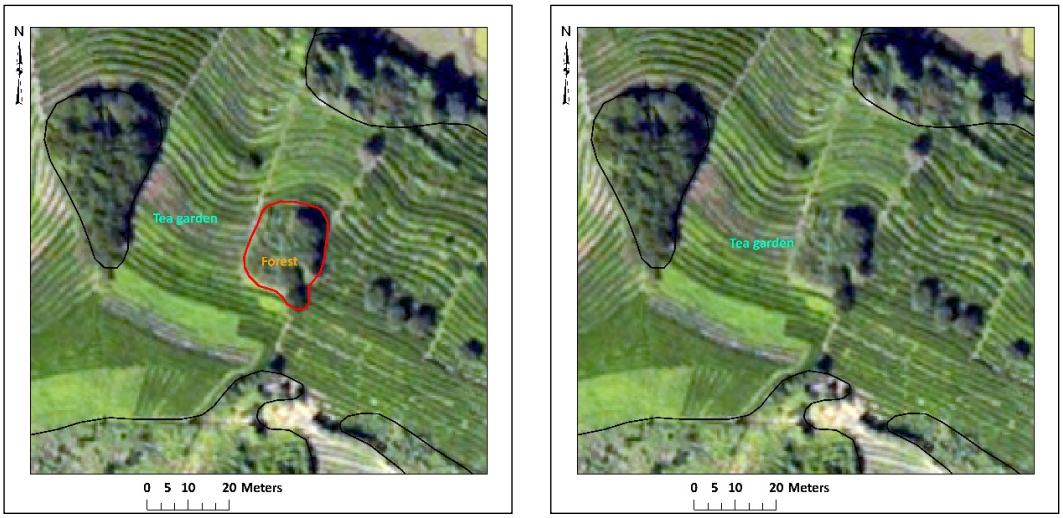
$$\begin{cases} p_{ij} = \dfrac{L_{ij}}{\sum_{i=1,i\neq j}^{n} L_{ij}}, i \neq j \\ p_{ij} = 0, \qquad i = j \end{cases}, \tag{4.7}$$

where L_{ij} is the total length of the shared boundary between the features of classes i and j, and n is the number of classes.

Apparently, L_{ij} equals L_{ji}. However, in the situation that a patch of class i is entirely enclosed by class j, the conversion from class j to i is not supposed to occur; thus, the shared boundary should not be counted into L_{ij}. An example is presented in Figure 4.5 to illustrate this situation. In such a situation, which is termed as the 'surrounded' situation in this study, only the one-way conversion can be considered. Thus, Equation (4.7) can be rewritten as

$$\begin{cases} p_{ij} = \dfrac{L_{ij} - \overline{L_{ij}}}{\sum_{i=1,i\neq j}^{n} (L_{ij} - \overline{L_{ij}})}, i \neq j \\ p_{ij} = 0, \qquad i = j \end{cases}, \tag{4.8}$$

where $\overline{L_{ij}}$ is the total boundary length of the features of class i that are enclosed by the features of class j.



(a) Before merging (b) After merging

Figure 4.5 Merging operation with a patch completed enclosed by another one. The shared boundary marked by a red line should not be counted for the conversion between class 'tea garden' to be merged into 'forest'.

Therefore, in view of the recovered omission errors and the resultant possibilities of conversions between land classes, the commission area of each LULC classes can be written in a matrix format, as shown as follows:

$$\begin{bmatrix} E_{\mathrm{c},1} \\ \vdots \\ E_{\mathrm{c},i} \\ \vdots \\ E_{\mathrm{c},n} \end{bmatrix} = \begin{bmatrix} p_{11} & \cdots & p_{1j} & \cdots & p_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ p_{i1} & \cdots & p_{ij} & \cdots & p_{in} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ p_{n1} & \cdots & p_{nj} & \cdots & p_{nn} \end{bmatrix} \begin{bmatrix} E_{\mathrm{o},1} \\ \vdots \\ E_{\mathrm{o},j} \\ \vdots \\ E_{\mathrm{o},n} \end{bmatrix}, \tag{4.9}$$

where $E_{\mathrm{c},i}$ is the total commission area for class i. Then, the actual area $A_{\mathrm{a},i}$ of class i can be calculated as

$$A_{\mathrm{a},i} = A_{\mathrm{m,i}} - E_{\mathrm{c},i} + E_{\mathrm{o},i}, \tag{4.10}$$

where $A_{\mathrm{m},i}$ is the measured area for class i.

On the basis of the results of omission and commission errors, a confusion matrix can be constructed, and the evaluation for the uncertainty brought by MMU becomes

possible. The confusion matrix is demonstrated in Table 4.3.

Table 4.3 Confusion matrix for uncertainty evaluation.

	C_1	C_2	⋯	C_n	Row total
C_1	$A_{\mathrm{m},1} - \sum_{j\neq 1}^{n}(p_{1j} \times E_{\mathrm{o},j})$	$p_{21} \times E_{\mathrm{o},1}$	⋯	$p_{n1} \times E_{\mathrm{o},1}$	$A_{\mathrm{a},1}$
C_2	$p_{12} \times E_{\mathrm{o},2}$	$A_{\mathrm{m},2} - \sum_{j\neq 2}^{n}(p_{1j} \times E_{\mathrm{o},j})$	⋯	$p_{n2} \times E_{\mathrm{o},2}$	$A_{\mathrm{a},2}$
⋮	⋮	⋮	⋱	⋮	⋮
C_n	$p_{1n} \times E_{\mathrm{o},n}$	$p_{2n} \times E_{\mathrm{o},n}$	⋯	$A_{\mathrm{m},n} - \sum_{j\neq n}^{n}(p_{nj} \times E_{\mathrm{o},j})$	$A_{\mathrm{a,n}}$
Column total	$A_{\mathrm{m},1}$	$A_{\mathrm{m},2}$	⋯	$A_{\mathrm{m},n}$	

4.4 Experiments and analysis

4.4.1 Data description

National Geographic Condition Monitoring is a nationwide land survey project in China. The classification system used for the land cover mapping has a hierarchical structure consisting of 10 level-1 classes, 46 level-2 classes and 77 level-3 classes. A typical land cover dataset in South China is selected from NGSM for the case study. This dataset consists of 117,717 land patches into 42 different land cover classes. The details for this dataset and the MMU for each land class are shown in Table 4.4.

Table 4.4 Details about the NGSM land cover dataset.

Index	Class code	Class level	Frequency	Total area (m^2)	MMU (m^2)	Description
1	**0110**	**Level 2**	**9,581**	**212,154,755**	**400**	**Paddy field**
2	**0120**	**Level 2**	**2,339**	**6,057,786**	**400**	**Dry land**
3	**0211**	**Level 3**	**11,676**	**136,625,583**	**400**	**Tree shrub garden**
4	0212	Level 3	71	550,098	400	Vine garden
5	0213	Level 3	7	6,247	400	Herbal garden
6	**0220**	**Level 2**	**16,221**	**103,089,107**	**400**	**Tea garden**
7	0250	Level 2	518	2,453,314	400	Nursery
8	0260	Level 3	43	79,518	400	Flower nursery

9	0291	Level 3	203	1,944,260	400	Other tree shrub field
10	**0311**	**Level 3**	**5,853**	**403,492,742**	**400**	**Broad-leaf forest**
11	**0312**	**Level 3**	**11,824**	**1,059,238,803**	**400**	**Coniferous forest**
12	0313	Level 3	93	2,599,415	400	Mixed forest
13	**0321**	**Level 3**	**5,289**	**34,591,101**	**400**	**Broad-leaf shrub**
14	**0340**	**Level 2**	**21,316**	**464,621,459**	**400**	**Bamboo forest**
15	0350	Level 2	41	674,008	1,600	Woodland
16	0360	Level 2	85	224,687	200	Green forest
17	**0370**	**Level 2**	**3,890**	**153,744,087**	**200**	**Artificial young forest**
18	**0411**	**Level 3**	**17,708**	**106,100,099**	**400**	**Highly covered grassland**
19	0422	Level 3	97	247,903	200	Green grass
20	0424	Level 3	404	3,186,694	400	Slope protection plant
21	0429	Level 3	61	315,939	400	Other artificial grassland
22	0511	Level 3	783	3993,353	1,600	Dense high buildings
23	**0521**	**Level 3**	**2,826**	**27,957,829**	**1,600**	**Dense low buildings**
24	0540	Level 2	594	267,466	200	Independent houses
25	**0550**	**Level 2**	**2,240**	**973,134**	**200**	**Independent low houses**
26	0601	Level 3	281	15,540,213	0	Road
27	**0710**	**Level 2**	**1,070**	**8,194,009**	**1,600**	**Hardened surface**
28	0721	Level 3	76	174,512	0	Dam
29	0750	Level 2	102	1,093,205	1,600	Greenhouse
30	0760	Level 2	25	55,072	400	Consolidated pool
31	0770	Level 2	29	373,606	1,600	Industrial facility
32	0790	Level 2	63	222,746	400	Other structures
33	0810	Level 2	58	1,296,753	1,600	Strip field
34	0821	Level 3	12	214,011	1,600	Tailings stack
35	0822	Level 3	1	4,181	1,600	Refuse stack
36	0829	Level 3	6	50,159	1,600	Other stack
37	0830	Level 2	117	3,755,181	1,600	Construction site
38	0890	Level 2	98	1,198,192	1,600	Other construction
39	0920	Level 2	327	2,222,819	1,600	Muddy surface
40	0940	Level 2	398	3,027,201	1,600	Gravelly surface
41	0950	Level 2	362	1,561,878	1,600	Rocky surface
42	**1001**	**Level 3**	**1,217**	**23,934,692**	**400**	**Water**

The original interpretation for this dataset is available, in which the land cover patches are collected with no limits on the patch size. This original interpretation is further used as the truth for the validation of the proposed method. In addition, Class 0601, which represents 'road', is excluded in the calculation of the conversion possibility, as discussed in the previous section.

In this case, only 14 classes with a frequency more than 1,000 are selected for

omission error modelling (bolded text in Table 4.4), and the remaining ones are reasonably expected to have minimal omission areas that can be ignored.

4.4.2 Experimental results

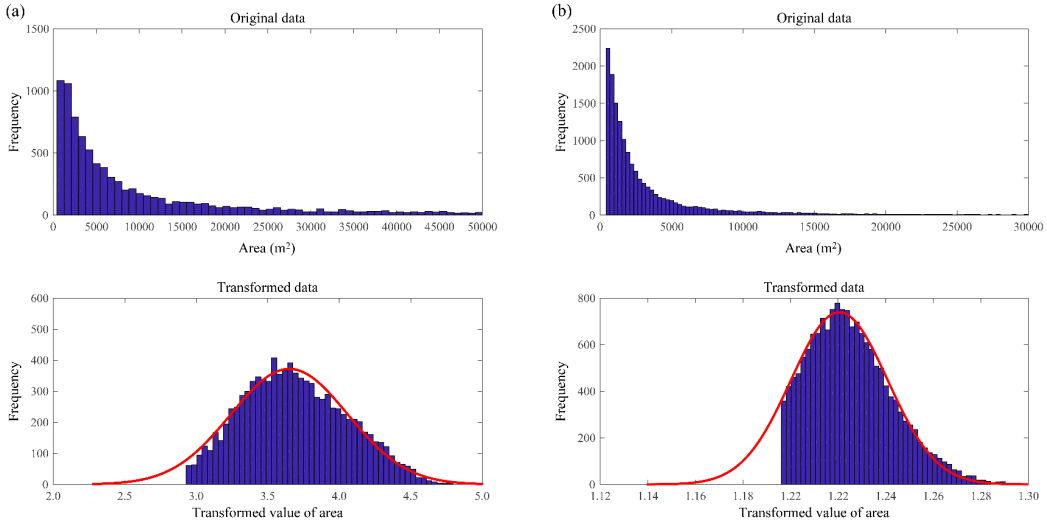


Figure 4.6 Two examples of the histograms of original and transformed data.

The histograms of two land classes are taken as examples to demonstrate the performance of the proposed method in recovering the omission errors. As shown in Figure 4.6, the original data of Classes 0110 and 0220 (i.e. 'paddy field' and 'tea garden') are observed to obey a typical positively skewed distribution. The transformations in Equation (4.3) are applied to the data to obtain multiple transformed data, and the normal function in Equation (4.4) is used to fit the transformed data. As shown in Figure 4.6, the transformed data, which have the least SSE in the fitting process, are likely to be normally distributed, whereas the blank areas under the left tail of the curves imply the potential existence of omitted land patches. Consequently, the omission errors can be recovered based on Equation (4.5), and the details for the recovery of each class are listed in Table 4.5.

Class 0521 (i.e. 'dense low buildings') is estimated to have the largest omission area of 1,838,879 m^2, which should be attributed to the discordance between the

relatively small size of buildings in reality and the specific large MMU (i.e. 1600 m^2). Class 0370 (i.e. 'artificial young forest') has the smallest omission area of 1,075 m^2, which might be explained by its relatively small MMU and low frequency. The results also demonstrate that the optimal transformation functions for land classes can be different, which reflects the adaptability of the proposed method on different original distributions.

Table 4.5 Fitting results of the selected classes. μ and σ are the estimated parameters for the transfer function with the least SSE.

Class	Description	μ	σ	Transformation	Predicted omission (m^2)
0110	Paddy field	3.650	0.409	$W = \ln(y)^{0.6}$	98,284
0120	Dry land	7.198	1.047	$W = \ln(y)$	86,051
0211	Tree shrub garden	1.527	0.044	$W = \ln(y)^{0.2}$	51,212
0220	Tea garden	1.221	0.020	$W = \ln(y)^{0.1}$	560,555
0311	Broad-leaf forest	1.884	0.123	$W = \ln(y)^{0.3}$	123,410
0312	Coniferous forest	0.186	0.062	$W = y^{-0.2}$	107,511
0321	Broad-leaf shrub	1.520	0.040	$W = \ln(y)^{0.2}$	21,995
0340	Bamboo forest	0.202	0.058	$W = y^{-0.2}$	269,230
0370	Artificial young forest	4.770	0.538	$W = \ln(y)^{0.7}$	1,075
0411	High covered grassland	2.268	0.137	$W = \ln(y)^{0.4}$	281,000
0521	Dense low buildings	2.718	0.291	$W = \ln(y)^{0.5}$	1,838,879
0550	Independent low houses	1.716	0.121	$W = y^{0.1}$	221,690
0710	Hardened surface	0.469	0.055	$W = y^{-0.1}$	682,777
1001	Water	3.839	0.710	$W = \ln(y)^{0.7}$	100,986

On the basis of the method introduced in Section 3.3, the results of commission errors are shown in Table 4.6. Notably, the commission area for Class 0601 (i.e. 'road') is 0 since this class has been excluded during the calculation, as discussed previously.

A Sankey diagram is drawn in Figure 4.7 to illustrate the relationship among omission errors, commission errors and the conversion possibility (Cuba 2015). The bars on the left side indicate the omission errors of each land class, whereas the one on the right side represents the commission errors. The height of a bar refers to the number of errors. The band connecting bars indicates the occurrence of conversion

between land classes, and the bandwidth represents the quantity of the corresponding conversion.

Table 4.6 Estimated commission errors.

Class	Description	Commission area (m^2)	Class	Description	Commission area (m^2)
0110	Paddy field	771,640	0511	Dense high buildings	139,493
0120	Dry land	222,371	0521	Dense low buildings	155,755
0211	Tree shrub garden	224,299	0540	Independent houses	19,429
0212	Vine garden	1,507	0550	Independent low houses	16,348
0213	Herbal garden	20	0601	Road	0
0220	Tea garden	238,308	0710	Hardened surface	146,669
0250	Nursery	13,574	0721	Dam	3,741
0260	Flower nursery	506	0750	Greenhouse	1,960
0291	Other tree shrub field	3,670	0760	Consolidated pool	1,253
0311	Broad-leaf forest	267,383	0770	Industrial facility	7,676
0312	Coniferous forest	718,172	0790	Other structures	5,299
0313	Mixed forest	3,378	0810	Strip field	4,577
0321	Broad-leaf shrub	115,963	0821	Tailings stack	1,430
0340	Bamboo forest	624,924	0822	Refuse stack	93
0350	Woodland	784	0829	Other stack	885
0360	Green forest	14,005	0830	Construction site	31,955
0370	Artificial young forest	92,882	0890	Other construction	10,809
0411	High covered grassland	421,963	0920	Muddy surface	10,272
0422	Green grass	11,606	0940	Gravelly surface	18,181
0424	Slope protection plant	21,071	0950	Rocky surface	1,204
0429	Other artificial grassland	1,897	1001	Water	91,667

As shown in Table 4.7, the most omission errors are caused by Class 0521 (i.e. ‘dense low buildings’), and the most commission errors are introduced by Classes 0110 and 0312 (i.e. ‘paddy field’ and ‘coniferous forest’). A large number of tiny patches of Class 0521 (i.e. ‘dense low buildings’) are estimated to be mapped as Class 0110 (i.e. ‘paddy field’). A significant conversion also exists from Class 0220 (‘tea garden’) on the left to Class 0312 (i.e. ‘coniferous forest’) on the right.

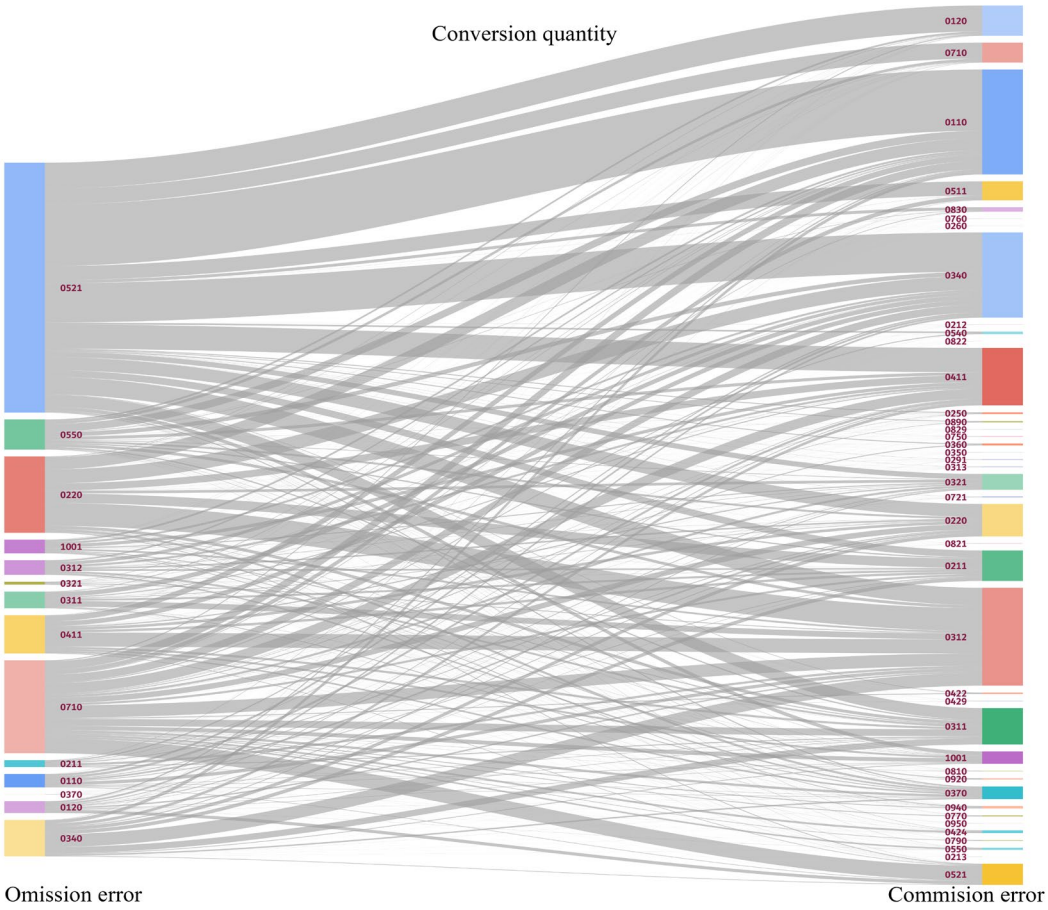


Figure 4.7 Sankey diagram of the conversion amongst classes.

Finally, the confusion matrix is constructed based on the expressions in Table 4.3. The producer's accuracy, user's accuracy, overall accuracy and kappa coefficient κ are listed in Table 4.7. The results show that Class 0550 (i.e. 'independent low houses') has the lowest producer's accuracy of 81.2%, whereas Class 0540 (i.e. 'independent houses') has the lowest user's accuracy of 92.7%. In contrast with the user's accuracy, the producer's accuracy is more sensitive to the manual operation because all the unselected classes will obtain the value 100%. Therefore, the user's accuracy should be a better indicator for the accuracy assessment. However, given the relatively smaller size of the omission areas compared with the total area, the overall accuracy and κ are high, which are 99.6% and 0.9980, respectively. This result shows the little effect of MMU on this dataset, which reflects the reasonability of the design of current MMUs.

Table 4.7 Results of accuracy assessment.

Class	Description	Producer's accuracy (%)	User's accuracy (%)	Class	Description	Producer's accuracy (%)	User's accuracy (%)
0110	Paddy field	100	99.6	0511	Dense high buildings	100	96.5
0120	Dry land	98.5	96.3	0521	Dense low buildings	93.8	99.4
0211	Tree shrub garden	100	96.2	0540	Independent houses	100	**92.7**
0212	Vine garden	100	99.7	0550	Independent low houses	**81.2**	98.3
0213	Herbal garden	100	99.7	0601	Road	100	100
0220	Tea garden	94.9	99.8	0710	Hardened surface	92.2	98.2
0250	Nursery	100	99.4	0721	Dam	100	97.9
0260	Flower nursery	100	99.4	0750	Greenhouse	100	99.8
0291	Other tree shrub field	100	99.8	0760	Consolidated pool	100	97.7
0311	Broad-leaf forest	100	99.9	0770	Industrial facility	100	97.9
0312	Coniferous forest	100	99.9	0790	Other structures	100	97.6
0313	Mixed forest	100	99.9	0810	Strip field	100	99.6
0321	Broad-leaf shrub	99.9	99.7	0821	Tailings stack	100	99.3
0340	Bamboo forest	99.9	99.9	0822	Refuse stack	100	97.8
0350	Woodland	100	99.9	0829	Other stack	100	98.2
0360	Green forest	100	93.8	0830	Construction site	100	99.1
0370	Artificial young forest	100	99.9	0890	Other construction	100	99.1
0411	High covered grassland	99.7	99.6	0920	Muddy surface	100	99.5
0422	Green grass	100	95.3	0940	Gravelly surface	100	99.4
0424	Slope protection plant	100	99.3	0950	Rocky surface	100	99.9
0429	Other artificial grassland	100	99.4	1001	Water	99.6	99.6
Overall accuracy (%)						99.6	
κ						0.9980	

4.4.3 Validation based on truth data

The relative accuracy r_i is used to measure the accuracy of each class i and can be expressed as follows:

$$r_i = 1 - \left| \frac{E_{o,i} - E_{c,i}}{A_{a,i}} \right|. \tag{4.11}$$

To validate the effectivity of the proposed method, as well as the proposed operation handling the 'surrounded' situation, the conversion possibility matrix is

computed again without considering the ‘surrounded’ situation. The corresponding outputs, together with the actual value of relative accuracy for each class, are compared with the evaluated results in the last section. Figure 4.8 shows the comparison results.

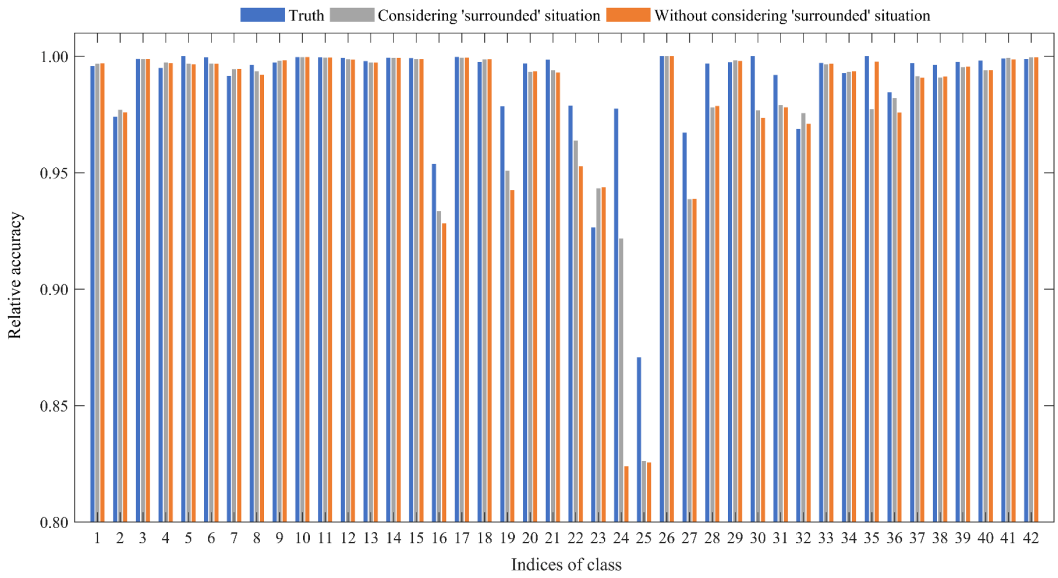


Figure 4.8 Relative accuracy for different models.

The relative accuracy is closer to the real value in most cases when ‘surrounded’ situation is considered. The most significant improvement occurs in the class of index 24, in which the relative accuracy increases from 0.820 to 0.925. Through our observation and analysis of the characteristics of those classes with similar improvement, we find that land classes that tend to have broken features are more likely to benefit from considering the ‘surrounded’ situation. Furthermore, the estimated relative accuracy for each class is consistently close to the actual value for most land classes, which improves the effectivity of the proposed method.

4.5 Summary

In this chapter, a reference-free method for evaluating the scale uncertainty brought by MMU in LULC data was proposed and demonstrated. An underlying assumption on the skew distribution of land patch size was suggested and verified with a real-life

dataset. Multiple transformation functions were designed to identify the potential skewed distribution of each land class, and a strategy based on curve fitting was used to select the optimal distribution parameters. The omission error caused by the MMU was recovered with the resultant distribution parameters, and the commission errors were subsequently computed based on the conversion possibilities amongst different land classes, which can be deduced from their adjacency relationship.

From the results of simulated and real data, the proposed method performs efficiently to reflect the general influence of the specific MMU on each land class. In addition to good performance, the method has some other merits. Firstly, the method utilises only the statistical information of data and does not require any reference data during the evaluation. Secondly, it strengthens the role of experimental data in the evaluation. Therefore, the evaluation becomes flexible and self-adaptive while the characteristics of different data vary.

Even though the proposed method is effective in both simulation and the case study, we should notice that the proposed method works well only if the assumption on positively skew distribution can be satisfied. Even though evidence from previous studies supports this assumption, reality can be more complex, resulting in irregulate data distribution especially for those man-made land objects with relative rigid sizes. To avoid the potential negative effect from such a situation, a quick validation can be conducted by applying a logarithm transformation on the raw data and manually checking the data distribution. In this way, the risk of irregular data distribution can be greatly reduced. Furthermore, even though this method currently assumes the commonest generalization rule and only consider the topological relationship to calculate the conversion probability between land classes, other information, such as the spectral information or semantical attributes of land patches, can be easily included with specific similarity models. These similarity measurements can be regarded as additional weights to improve the conversion probability generated by current generation rule, and thus producing more realistic estimations of commission errors.

Chapter 5 Adaptive uncertainty models for trajectory big data

In this chapter, we focus on human activity data, which is another main category SBD, as discussed before.

Amongst the various types of human activity data, trajectory big data are of great importance. These data directly capture the spatiotemporal movements and flows of target objects. Generally, trajectory data consist of finite sequences of spatiotemporal points. However, given the inevitable measurement error and the limited representability of finite points for continuous human movement, the uncertainties in trajectory data are not negligible. This chapter introduces a novel position uncertainty model for trajectory data called adaptive error ellipse (AEE) model, which generates the most appropriate error ellipse based on the global characteristics of trajectories at a given expected accuracy level. In addition, a broad AEE (BAEE) model is developed in approximating the uncertainty area between two consecutive trajectory points to handle the measurement errors in observations, such as GPS errors. The effectiveness of the proposed models is demonstrated through experiments on five real-life datasets and a case study by comparing with state-of-the-art methods.

5.1 Trajectory uncertainty models

5.1.1 Overview of trajectory uncertainty models

Object movement is always represented by the trajectory data. A raw trajectory comprises a finite number of sequence points $<pt_1, \dots, pt_n>$, where $pt = (x, y, t)$, x and y are the geographic coordinates of the target object at time t.

Generally, the uncertainty of trajectory data originates from two types of error, namely, measurement and sample errors. Measurement error, such as GPS error, is known to prevalently exist in the positioning process due to the limited accuracy of positioning technology, which may differ from different positioning modes, devices and algorithms. Sample error is caused by the finite number of discrete points in a raw

trajectory, which cannot fully capture the continuous movement in reality (Jeung et al. 2014). Given the existence of measurement and sample errors, the uncertainty of trajectory data, in practice, is manifested as the positions of sampled points being inaccurate and the locations of the target object between two sampled points being unknown.

Various methods have been proposed to reconstruct the actual motion from the finite sampled points. For cases that emphasise the moving flow rather than the actual movements (e.g. origin-destination (O–D) analysis), interpolation algorithms are commonly applied to recover the continuous movements based on sampled points in the literature (Hoteit et al. 2014). however, these interpolation algorithms easily fail to represent the actual movement because they assume that the motion between two sampled points is close to a specific curve (e.g. straight line for linear interpolation), which is greatly questionable due to the complicated mobility pattern and environmental context in the real world (Kuijpers and Othman 2006). Moreover, the effectiveness of interpolation algorithms significantly declines on low-sample trajectories because low sample rate leaves relatively long-time actual movement unknown, making the interpolation results unreliable (Ranu et al. 2015).

In the light of the limitations of interpolation algorithms, researchers have turned to construct uncertainty models rather than a single interpolated curve to cover as many potential locations of the target object as possible. In that sense, the reconstructed trajectory is represented by several uncertain regions that are also called uncertain trajectory. Conventional trajectory uncertainty models include the beads model (Hornsby and Egenhofer 2002) and the cylinder model (Pfoser and Jensen 1999, Trajcevski et al. 2004, Kuijpers and Othman 2009). In view of the restriction of road networks on the movement of particular objects, such as cars and buses, some network-constrained models have been established (Zheng *et al.* 2011, Chen, Tang, *et al.* 2015). These uncertainty models have been widely used for efficient query processing (Trajcevski *et al.* 2010, Niedermayer *et al.* 2013a), similarity analysis

(Niedermayer *et al.* 2013b, Furtado *et al.* 2018), uncertainty visualisation (Huang and Wong 2015) and sampling strategy (Ranacher and Rousell 2013).

The cylinder model is established by buffering the line segments between consecutive sampled points, where the buffer size is predefined based on prior knowledge (Trajcevski et al. 2004, Bonchi et al. 2011). Similar to the linear interpolation algorithm, this model assumes the linear motion of an object between consecutive sampled points while allowing an uncertain range (i.e. the buffer) for the nonlinear or even random movements. In 3D *x*–*y*–*t* space, these buffered line segments become sheared cylinders to represent the uncertain trajectory. The cylinder model, however, highly relies on the line segments between sampled points, which inherits the drawbacks of interpolation methods, as mentioned above. Moreover, the difficulty in selecting an appropriate buffer size also hampers the robustness of the cylinder model in practice.

The development of the beads model, also called space-time prism model in some literature, can be traced back to Pfoser and Jensen's work (1999), in which they considered the fact that an object's moving distance is restricted by the maximum speed. Therefore, using two consecutive sampled points pt_1, pt_2 as the foci and the maximum moving distance d during time $|t_{pt_1} - t_{pt_2}|$ as the major axis, an error ellipse was adopted to represent the upper bound of the potential locations that the target object may visit between two consecutive sampled points. This ellipse evolves into a space–time prism in the 3D *x*-*y*-*t* space (Kuijpers and Othman 2009). However, the error ellipse (or space-time prism) model always produces an oversize uncertain region since the actual moving speed is easily less than the theoretical maximum. This limitation also makes the beads or ellipses extremely large on low-sampled trajectory datasets, thereby significantly reducing the practical value of the beads model (Jeung et al. 2014).

Amongst these aforementioned models, network-constrained methods can provide a relatively tight uncertain range to capture the uncertain trajectory with

additional network information (Jeung et al. 2014). However, it has no advantages for trajectories in free space, such as the vessel trajectory on the sea, and loses its efficiency when road information is unavailable.

In sum, current uncertainty models are limited by intuitive parameter setting and external information to different extents. However, from the robustness point of view, the beads model seems to be superior to others because it relies on a relatively realistic assumption and requires no additional information. Furthermore, the beads model is universal because it is suitable for either free-space or constrained movements. Therefore, we take the beads model as the basis and attempt to enhance its robustness and effectiveness in practice.

5.1.2 Improved beads model and its pitfalls

An overestimated speed often produces unrealistically large uncertain regions that significantly influence the reliability of the beads model. To address the drawbacks of the traditional beads model, Furtado et al. (2018) introduced a distance metric named approximate upper bound (AUB) to dynamically determine the length of the long axis of each error ellipse. In view of the influence of the direction of the coordinate system on the Manhattan distance metric, the AUB method uses the maximum Manhattan distance as the upper bound distance to generate the final error ellipse. Mathematically, the AUB distance between two spatial points pt_1 and pt_2 presents a constant proportion to the ED, which is expressed as (Furtado et al. 2018)

$$\mathrm{AUB}(pt_1, pt_2) = \sqrt{2} * \mathrm{Euclidean}(pt_1, pt_2). \tag{5.1}$$

The constant $\sqrt{2}$ refers to the largest ratio between the Manhattan distance and ED when the arc tangent between these two points is 45°. The constant is used to reduce the influence from the direction of the coordinate system on Manhattan distance measurements. On the basis of the experimental results in the original paper (Furtado et al. 2018), the AUB method is more efficient than the traditional beads

model in 2D space, which significantly reduces the total area of error ellipse while covering most actual motion points.

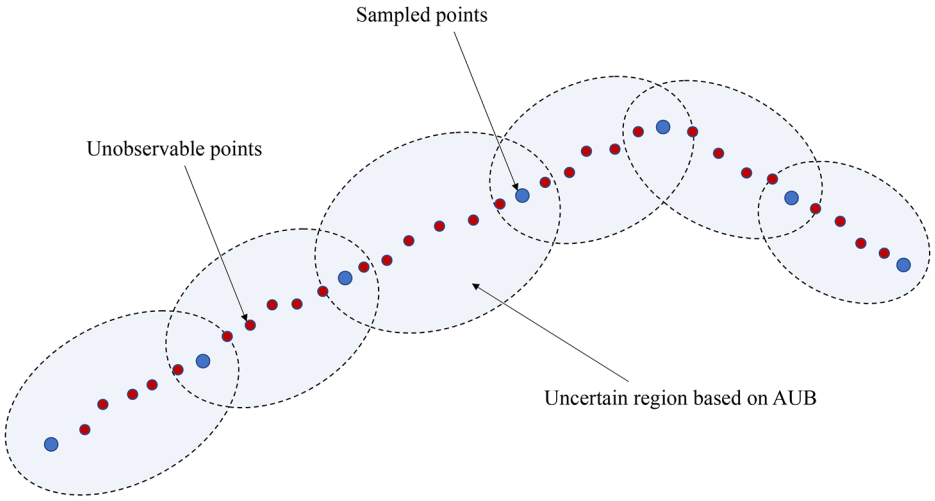


Figure 5.1 Uncertain regions for a linear movement based on AUB method.

The underlying assumption of AUB method is that an object's actual movement between two sampled points, which can be taken as a pair of O–D points, is constrained by an underlying network, and the network distance (e.g. road distance) can be approximated by the Manhattan distance. This assumption is valid for specific situations, such as the movement of cars within a planned city traffic network such as the one in Manhattan. However, the literature indicates that Manhattan distance, as well as ED, is inadequate for spatial distance measurement considering the complex situations in reality (Shahid et al. 2009, Bora et al. 2014). As a result, the AUB method can still overestimate the uncertain region due to its relatively fixed ellipse size, which is entirely determined by the distance between two sampled points. A simple example is given in Figure 5.1 to illustrate the overestimated situation. From the figure, the object almost moves in a straight manner; thus, almost all the unobservable points are located near the line segments connecting the sampled points. However, if the line segments are extremely long, then the AUB method will consequently output enormous error ellipses, of which a high proportion of uncertain region has little practical value.

5.2 Construction of AEE model

5.2.1 Introduction on Minkowski distance metric

The Manhattan distance or ED metrics cannot reflect the actual moving distance in various environments. In this respect, the Minkowski distance metric can be used because it can account for different non-ED metrics by varying its exponent parameter p (Lu et al. 2016). Given two 2D spatial points $pt_1 = (x_1, y_1)$ and $pt_2 = (x_2, y_2)$, the Minkowski distance metric is expressed as:

$$MD_p(pt_1, pt_2) = \left(|x_1 - x_2|^p + |y_1 - y_2|^p \right)^{1/p}. \tag{5.2}$$

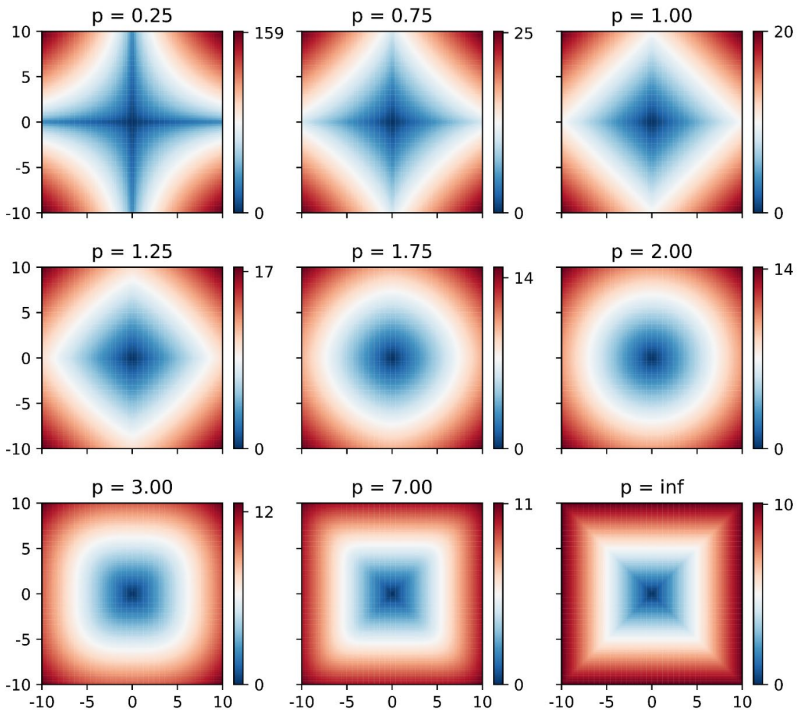


Figure 5.2 Surface plot of the Minkowski distance metric with different p values.

The performance of the Minkowski distance metric is directly determined by the value of p. For example, when p is $1, 2$ or ∞, the Minkowski distance becomes the Manhattan, Euclidean and Chebyshev distances in a Euclidean space. Taking the 2D *x*–*y* space as an example, we plot the distance surfaces of different values of *p* in Figure 5.2. From the figure, different values of *p* provide different spatial distance metrics,

and close p values lead to similar patterns.

5.2.2 Building AEEs based on Minkowski distance metric

With rich variability, the Minkowski distance metric provides a more flexible selection of distance metric rather than the rigid use of the Manhattan distance metric in AUB methods. For this reason, by extending the basic idea of AUB method, the proposed AEE model adopts the Minkowski distance to improve its robustness.

In the 2D Euclidean space, the actual moving distance between two sampled points should always be larger than the corresponding ED (Furtado et al. 2018). Therefore, representing the actual moving distance by Minkowski distance metric implies that the value of p should not be over 2. However, the measurement of Minkowski distance differs with respect to different directions of the coordinate system. Similar to the process of obtaining the maximum Manhattan distance in AUB method, we should find the maximum Minkowski distance as the upper bound distance by rotating the coordinate system to construct the error ellipse.

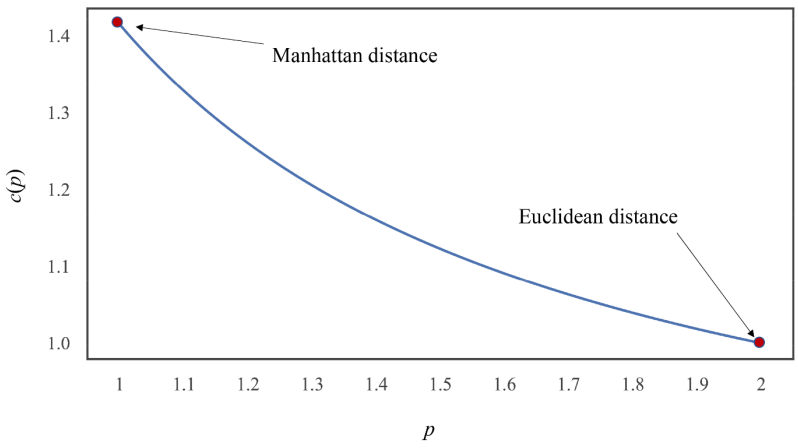


Figure 5.3 Ratio between maximum Minkowski distance and ED with different p values.

Let A denote an arbitrary point (x, y) and C denote the circle centring at the origin point O with a radius of $\mathrm{Euclidean}(AO)$. The Minkowski distance between AO is expressed as

$$MD_p(AO) = \left(|x|^p + |y|^p \right)^{1/p}. \tag{5.3}$$

With the condition $x^2 + y^2 = r^2$ and $p \leq 2$, $MD(AO)$ reaches the maximum when $|x| = |y|$. The ratio $c(p)$ between Euclidean(AO) and maximum $MD(AO)$ regarding different p values is plotted in Figure 5.3. Notably, the ratio between the maximum Minkowski distance and ED dramatically declines as p increases. In that sense, as our main objective is to reduce the redundant uncertain region in AUB method, we expect the error ellipse in our model to be no larger than that in the AUB method. Accordingly, we control the value of p to be no less than 1. Therefore, given two sampled points $pt_1 = (x_1, y_1)$ and $pt_2 = (x_2, y_2)$, the major axis for AEE model is defined as

$$AEE(pt_1, pt_2) = c(p) * \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}. \tag{5.4}$$

5.2.3 Selecting the optimal parameter for Minkowski distance metric

The value of parameter p is crucial for the performance of the Minkowski distance metric. To seek the optimal parameter p, we introduce an optimisation process based on the characteristics of the raw trajectory. The details of the optimisation process are demonstrated in Algorithm 1.

The main idea of Algorithm 1 is to test different p values on the raw trajectory and determine the one with the largest accuracy. A threshold ACC is predefined to control the accuracy of the optimisation process, and Δp is defined as the incremental step for testing the p values ranging from 1 to 2. In Algorithm 1, each pair of temporally adjacent points p_k, p_{k+1}, p_{k+2} forms a subunit. With the given p value and corresponding ratio c, the length of major and minor axes is obtained based on Equation (4), and an ellipse is generated between p_k and p_{k+2} through the function *BuildEllipse*(). Then, function *checkInside*() is used to check if p_{k+1} falls in the ellipse. By going through all pairs of points, the array *InsiderNum* is used to record the total number of points that falls in the corresponding ellipse. The final accuracy

for the current p value is computed as the proportion between *InsiderNum* and the total number of point pairs (i.e. $n-2$). After all the selected p values are tested, we filter the largest qualified p, which corresponds to an accuracy higher than the threshold ACC, as the output. If a qualified p does not exist, then the output is set to the largest p value corresponding to the maximum accuracy.

Algorithm 1 Optimizing the Minkowski distance metric parameter p for a given raw trajectory

Input: A raw trajectory with n sampled points $RT =< pt_1, , pt_n >$, a threshold of expected accuracy ACC, an incremental step Δp

Output: The adaptive parameter $p(RT)$

```
k = 0, p = 1, res = []; //initialization
while p <= 2 do
    c=Ratio(p); //obtain the ratio with respect to p
    for k = 0 : n − 2 do
        MajorAxesLen=r*Eud(pt_k, pt_{k+2}); //compute the length of major axes
        MinorAxesLen=(c^2 − 1)^{1/2}*Eud(pt_k, pt_{k+2}); //compute the length of minor axes
        Ellipse=BuildEllipse(MajorAxesLen,MinorAxesLen); //generate the error ellipse between pt_k and pt_{k+2}
        IsInside=checkInside(pt_{k+1},Ellipse); //check if pt_{k+1} is inside the ellipse
        if IsInside then
            InsideNum[p] + +; //record the number of inside points
        end if
    end for
    res[p]=InsideNum[p]/(n − 2); //compute the accuracy results regarding current p value
    p = p + Δp;
end while
QualifiedP=GetQualifiedValues(res > ACC); //get the p values that satifies the limitation on accuracy
if QualifiedP == Null then
    p(RT)=max(p) s.t. res[p] ==max(res)
else
    p(RT)=max(QualifiedP); //use the largest p as the final result
end if
Return p(RT);
```

In Algorithm 1, the movement pattern is extracted from the raw trajectory, and no additional information is required. This condition significantly increases the robustness of this algorithm because obtaining detailed road network information is practically always difficult. Furthermore, the accuracy of the resultant model can be roughly estimated and controlled by adopting the accuracy threshold to filter qualified p values. This merit endows the AEE model to be significantly more advantageous than the AUB model because the performance of the latter is unpredictable and uncontrolled, especially when the optimal p is considerably larger than 1 (i.e. the overestimation in AUB method is expected to be serious). Furthermore, the optimal p

is selected as large as possible in each situation (whether qualified values), which maximises the reduction of the redundant uncertain regions.

The AEE model is theoretically more robust than the AUB method due to the variability of the Minkowski distance metric. However, a limitation for the AEE model is that the measurement error in raw trajectories is not considered. Therefore, in the following section, we further introduce the BAEE model, which considers the measurement error in each sampled point during the error ellipse construction.

5.3 Construction of BAEE model

5.3.1 Foundation of BAEE model

The basic idea for the BAEE model is that the sampled points in a raw trajectory are not precisely where the object travels in reality. Given the pervasive existence of measurement error, such as GPS error or error from other positioning techniques, an uncertain circle centring is always used to reflect the maximum position offset from the actual location to the sampled point in Euclidean space (Epple 2006, Zhang and Hsu 2019). Accordingly, we consider the uncertain circle when constructing the error ellipse and thus establish a BAEE model.

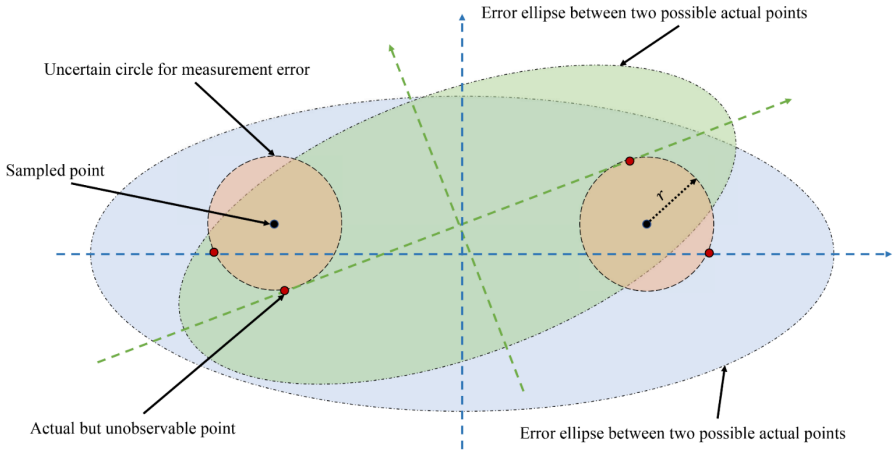


Figure 5.4 Foundation of BAEE model.

The foundation of the BAEE model is illustrated in Figure 5.4. Different from the

AEE model, the foci of each error ellipse are no longer the two sampled points but any two arbitrary points within the error circle centring at the sampled points with a radius of r. Therefore, the infinite error ellipses based on different pairs of arbitrary points will cover a relatively broader region than the single ellipse in AEE model with the same p value. We call this broader region the BAEE model.

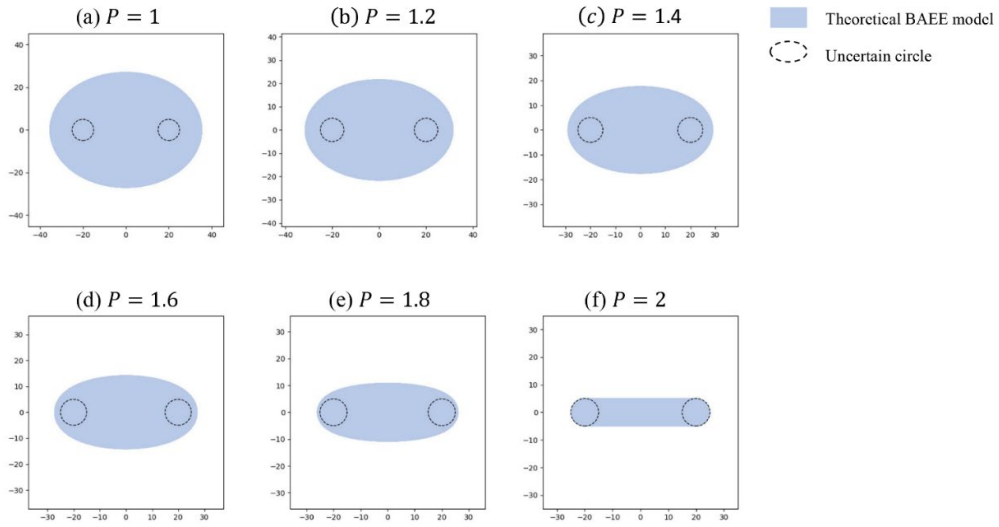


Figure 5.5 Theoretical BAEE model (blue region) for different p values. The measurement error (i.e. the radius of the black circle) is set to 5. The coordinates for two foci are (−10, 0) and (10, 0).

Through intensive computation, we can obtain a theoretical BAEE model for visualisation. Theoretically, a standard formula for BAEE model is unavailable. This phenomenon can be simply proven by considering an extreme situation shown in Figure 5.5(f), where p is 2. In this manner, all the supposed error ellipses become straight lines, and a simulated region of a combined shape composed of a rectangle and two semi-circles is generated. Figure 5.5 (a)Figure 5.5(e) show that in almost all cases, the theoretical BAEE model appears to be an ‘ellipse’.

5.3.2 Approximated ellipse for BAEE model

The computation of the theoretical BAEE model is intensive and seriously affects the efficiency of further analysis. Therefore, a shape with a standard formula should be

determined to approximate the theoretical uncertain region. On the basis of Figure 5.5 and preliminary tests, an ellipse is selected for the approximation. The ellipse should satisfy two conditions: 1) the major axis of the ellipse should be equal to the longest radial distance along the direction of the line between two sampled points; and 2) the minor axis of the ellipse should be equal to the longest radial distance along the vertical direction of the line between two sampled points. Therefore, the approximated ellipse can be determined by computing the extreme value of the theoretical BAEE model.

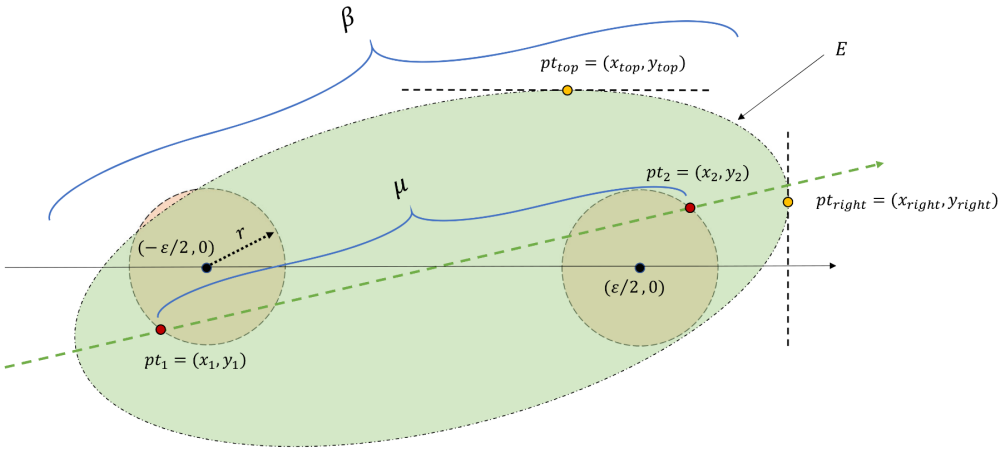


Figure 5.6 Diagram for the ellipse generated by two arbitrary points.

Figure 5.6 shows a diagram to help illustrate the computation process for the extreme values of the BAEE model. Let $pt_1 = (x_1, y_1)$ and $pt_2 = (x_2, y_2)$ denote the two arbitrary points on the uncertain circles centring at two sampled points in a Euclidean space. To simplify the computation process, the y-axis coordinates of the sampled points are set to zero, and the x-axis coordinates are set to $-\varepsilon/2$ and $\varepsilon/2$. The arbitrary ellipse E between pt_1 and pt_2 can be expressed as

$$\sqrt{(x-x_1)^2+(y-y_1)^2}+\sqrt{(x-x_2)^2+(y-y_2)^2}=\beta\,, \tag{5.5}$$

where β denotes the major axis of the arbitrary ellipse E. Given a variable p for Minkowski distance measurement, β is proportional to the ED between pt_1 and pt_2, which can be expressed as

$$\beta = Ratio(p) * \mu = c * \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} . \tag{5.6}$$

Therefore,

$$\frac{\sqrt{(x-x_1)^2+(y-y_1)^2}+\sqrt{(x-x_2)^2+(y-y_2)^2}}{c*\sqrt{(x_1-x_2)^2+(y_1-y_2)^2}}=1 . \tag{5.7}$$

Suppose the measurement error is r. pt_1 and pt_2 should satisfy the restriction from the uncertain circle. Therefore, any point (x, y) on the boundary on the arbitrary ellipse E should satisfy

$$\begin{cases} \dfrac{\sqrt{(x-x_1)^2+(y-y_1)^2}+\sqrt{(x-x_2)^2+(y-y_2)^2}}{c*\sqrt{(x_1-x_2)^2+(y_1-y_2)^2}}=1 \\ (x_1+\dfrac{\varepsilon}{2})^2 + {y_1}^2 = r^2 \\ (x_2-\dfrac{\varepsilon}{2})^2 + {y_2}^2 = r^2 \end{cases} . \tag{5.8}$$

As shown in Figure 5.6, each arbitrary ellipse E has its own extreme points, such as $p_{top} = (x_{top}, y_{top})$ and $p_{right} = (x_{right}, y_{right})$. Taking the top point $(\mathrm{X}_{top}, Y_{top})$ and right point $(\mathrm{X}_{right}, Y_{right})$ of the BAEE model as an example yields

$$\begin{cases} Y_{top} = \max\limits_{E\in\Omega}(y_{top}(E)) \\ X_{\text{right}} = \max\limits_{E\in\Omega}(x_{right}(E)) \end{cases}, \tag{5.9}$$

where Ω denotes the infinite set of all possible arbitrary ellipses. To obtain these two points, we initially deduce the formula for y_{top} and x_{right} based on Equation (5.8). Then, we compute their maximum by changing the locations of pt_1 and pt_2. The detailed derivation process is described in Appendix A. Finally, we have

$$\begin{aligned} (X_{right}, Y_{right}) &= (c*(\frac{\varepsilon}{2}+r), 0) \\ (X_{top}, Y_{top}) &= (0, r*c+\frac{\varepsilon}{2}*\sqrt{c^2-1}) \end{aligned} . \tag{5.10}$$

On the basis of Equation (5.10) and the symmetrical characteristics of the BAEE model, we can easily generate an ellipse to approximate the theoretical BAEE model, and this ellipse enables fast inside-point checking and parameter optimisation, which is similar to the process in the AEE model.

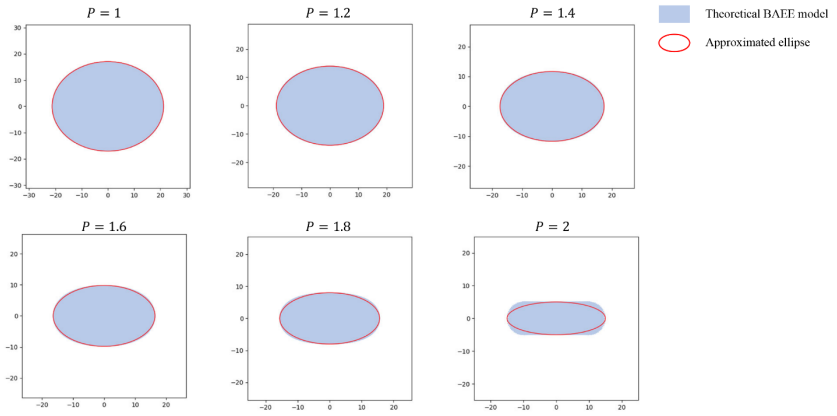


Figure 5.7 Example of the comparison between theoretical BAEE models and approximated ellipses. The radius of the uncertain circle (i.e. measurement error) is 5. The coordinates for two foci are (−10, 0) and (10, 0).

To show the effectiveness of the approximation, a visual comparison is made between the theoretical BAEE model and the corresponding approximated ellipse in Figure 5.7. For a relatively small p (e.g. $p < 1.4$), the ellipse fits well with the theoretical model. As p increases to 2, the difference gradually becomes distinguished around the ellipse. However, the theoretical model seems to be only slightly larger than the approximated ellipse, except for the situation with p being close to 2. Given the significant improvement in computation efficiency, we deem such differences negligible.

To further verify the effectiveness of the proposed approximation process, we provide some examples to illustrate the difference between theoretical BAEE and approximated ellipse. The difference is related to three parameters, that is, foci distance ε, measurement error r and Minkowski coefficient p. As shown in Figure 5.7, distinguishable difference appears when p is close to 2. Therefore, we fix $p = 1.9$ and compare the theoretical BAEE and approximated ellipse in three different

situations, that is, $\varepsilon \gg r$, $\varepsilon = r$ and $\varepsilon < r$. For different values of ε and r, the difference seems to be negligible because it is relatively smaller compared with the theoretical model. Therefore, we can safely use the approximated ellipse instead of theoretical BAEE in practice.

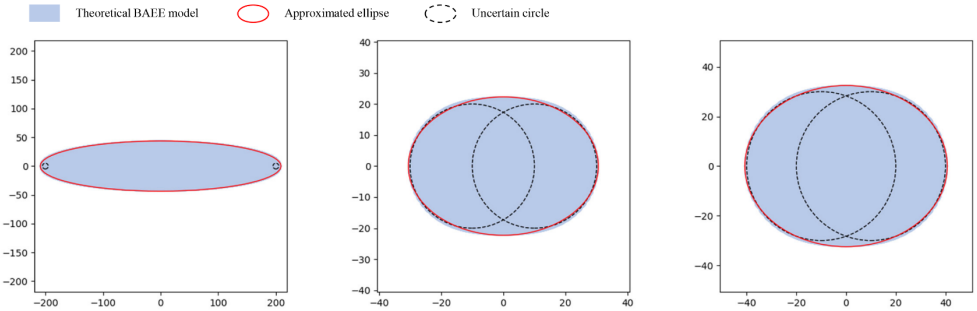


Figure 5.8 Effect from different parameters on the approximated ellipse.

Notably, all the ellipses constructed above are hypothetical, and the sampled points are located at two axisymmetric points on the x-axis. The sampled points in the real case are located differently in a specific coordinate system. Thus, the corresponding ellipse can be constructed by rotating and offsetting the hypothetical ellipse, of which the process is no longer discussed here. A useful reference providing detailed guidance can be found in https://www.visiondummy.com/2014/04/geometric-interpretation-covariance-matrix/.

5.4 Experiments and analysis

5.4.1 Data description and experimental setup

We evaluate the proposed models with five real-life trajectory datasets with different mobility patterns and sampling rates. Short descriptions of these five datasets are given as follows.

- Geolife[1] (Zheng et al. 2009): The dataset comprises outdoor movements of 182 users over five years, and the trajectories of 72 users are labelled with transportation mode (e.g. walk, bike and bus). The trajectories are collected by different GPS loggers and GPS phones. The sampling rate generally varies from 1 s to 5 s.
- Indoor[2] (Brščić et al. 2013): This dataset comprises visitors' movement records in the ATC shopping centre covering an area of approximately 900 m^2 in Osaka. The sampling rate for the trajectories ranges from 0.03 s to 0.06 s.
- Cab A[3]: This dataset is retrieved from Didi Chuxing, which provides ride-hailing in China. The dataset contains one-month cab trajectories in Xi'an. The trajectory points are sampled every 2–4 s. To capture the riding activities, the trajectories are labelled with additional riding information, such as the driver's identity and travel order's identity.
- Cab B: This dataset is also retrieved from Didi Chuxing. It comprises one-month cab movements in Chengdu. The sample rate also ranges from 2 s to 4 s, and anonymised driver and order information are attached.
- Bus[4] (Dias and Costa 2018): This dataset contains real-time position data of more than 12,000 buses for one month from the city of Rio. The bus movements are sampled strictly every minute, and some transport data are also provided, including bus line, bus ID and speed.

Given the enormous data volume of the complete trajectories in Indoor, Cab A, Cab B and Bus datasets, which is caused by the massive number of travellers or

[1] http://research.microsoft.com/en-us/projects/ geolife/

[2] http://www.irc.atr.jp/crest2010_HRI/ATC_dataset/

[3] https://gaia.didichuxing.com

[4] https://crawdad.org/coppe-ufrj/RioBuses/20180319/

vehicles, we randomly select one day for these datasets and retrieve the corresponding trajectories for our experiments. For the Geolife dataset, all trajectories are used. In addition, we apply a time threshold for the gap between two consecutive points for each dataset to avoid the potential uncertainty brought and identify meaningful consequently movements. On the basis of the characteristics of trajectories and the sampling rate in each dataset, the time threshold is set to 5 min for the Geolife, Cab A, Cab B and Bus datasets, and 5 s for the Indoor dataset.

Table 5.1 Statistics of the experimental data form five real-life datasets.

Dataset	Tested trajectories	Tested points	Average distance	Average speed	Sampling rate
Geolife	6,758	6,990,847	11.64 m	9.57 m/s	< 2 s
Indoor	17,380	16,762,441	0.05 m	1.14 m/s	< 0.05 s
Cab A	105,885	27,510,263	16.95 m	5.30 m/s	< 4 s
Cab B	213,887	41,574,989	19.84 m	6.15 m/s	< 4 s
Bus	82,868	3,675,496	251.27 m	4.68 m/s	< 70 s

The real movements are theoretically required for an ideal analysis in our experiment. However, actual movements are not approachable in practice. To address this problem, we downsample the raw trajectories, adopt the downsample trajectories to generate error ellipses and compare with the raw trajectories. To make the above process meaningful, low sampled trajectories, which indicate the trajectories with high sampling rate, should be filtered. Therefore, in this experiment, we select only high-sampled trajectories that have an average sampling rate of lower than 2 s for Geolife, 0.05 s for Indoor, 4 s for Cabs A and B and 70 s for Bus. The statistics of the qualified data from each dataset are reported in Table 5.1. In view of the computation cost, the accuracy threshold ACC is set to 95%, and the incremental step Δp is set to 0.05.

5.4.2 Accuracy of AEE and BAEE

To evaluate the proposed methods, we perform a comparison with the state-of-the-art method, AUB model. This part of the experiment consists of three steps: 1) downsampling the raw trajectories with a given sampling rate, 2) creating error ellipses

for all the consecutive points in the downsampled trajectories and 3) computing the average area coverage of these ellipses and the proportion of the points in raw data that have been covered by these ellipses.

We exemplify the general performance of the proposed methods with selected downsampling rate for each dataset. The downsampling rate is generally set as 15 times the lowest average sampling rate for each dataset. Finally, the downsampling rate is 30 s for Geolife, 60 s for Cabs A and B, 750 ms for Indoor and 900 s for Bus. To build the BAEE model, the measurement error is set to 20 m for Geolife, Cabs A and B, and Bus datasets; this value is typically used in related studies (Lou *et al.* 2009, Liu, Biagioni, *et al.* 2012). The measurement error for Indoor dataset is set to 100 mm, which is approximately the average level reported in the original paper (Brščić et al. 2013).

For a robust comparison, we take the area coverage of the AUB model as the baseline area and calculate the median ratio of the reduced area coverage brought by AEE and BAEE models. The baseline AUB model for validating BAEE also includes the measurement error for a fair and meaningful comparison, which is named B-AUB in the following discussion. In addition, the average proportion of raw trajectory points coverage is also recorded for each model to indicate model accuracy. We also use the accuracy per unit area to reflect the reliability of uncertain regions, where the unit area is defined as the area of the baseline model (i.e., AUB or B-AUB). Mathematically, the model reliability for a dataset of N trajectories is expressed as

$$R=\frac{1}{N}\sum_{i}^{N} Acc_i * A_i / UA_i \tag{5.11}$$

where Acc is the proportion of raw trajectory points covered by the tested model and A is the corresponding model area; UA is the area of the baseline model (i.e., AUB or B-AUB). The comparison results are summarised in Table 5.2 and Table 5.3.

The AEE and BAEE models can significantly reduce the area of uncertain regions,

which reaches the maximum 36.60% for the AEE model in Cab B and 49.75% for the BAEE model in Geolife. However, they exhibit comparative performance in covering a high proportion of raw trajectory points as AUB and B-AUB models. Particularly, for the Bus dataset, the decreased uncertain region area is significantly smaller than those for other datasets. Through our investigation and analysis, such small decrement is caused by the relatively low sampling frequency, which makes the movement between two sampled points less predictable and thus requiring a broader uncertain region. When the measurement error in sampled points is considered, the BAEE model reduces a higher proportion of uncertain region compared with the performance of AEE, in which the measurement error is neglected.

The reliability of the proposed model is significantly larger than the that of AUB or B-AUB in all cases, especially for the cases of Indoor, Cab A and Cab B, where the reliability are more than twice than in terms of the reliability of the corresponding baseline model (i.e., AUB or B-AUB). This finding proves the effectiveness of the AEE and BAEE in enhancing the reliability of uncertain regions, which also agrees to our expectation that AEE and BAEE can reduce the overestimated uncertain regions in the AUB method.

Table 5.2 Performance of AEE model with specific parameters on five datasets.

	Average ratio of reduced area (%)				
	Geolife	Indoor	Cab A	Cab B	Bus
AUB (baseline)			-		
AEE	23.07	32.33	34.82	**36.60**	8.22
	Average proportion of raw trajectory points coverage (%)				
	Geolife	Indoor	Cab A	Cab B	Bus
AUB	96.69	89.80	98.57	98.41	90.51
AEE	95.66	87.56	95.76	94.93	89.52
	Average reliability				
	Geolife	Indoor	Cab A	Cab B	Bus
AUB	0.97	0.90	0.99	0.98	0.91
AEE	1.63	2.36	**2.38**	2.29	1.36

Table 5.3 Performance of BAEE model with specific parameters on five datasets.

	Average ratio of reduced area (%)				
	Geolife	Indoor	Cab A	Cab B	Bus
B-AUB (baseline)			-		
BAEE	**49.75**	48.95	40.46	41.44	12.68
	Average proportion of raw trajectory points coverage (%)				
	Geolife	Indoor	Cab A	Cab B	Bus
B-AUB	99.81	99.96	99.42	99.39	94.88
BAEE	99.28	98.92	97.55	96.95	93.72
	Average reliability				
	Geolife	Indoor	Cab A	Cab B	Bus
B-AUB	1.00	1.00	0.99	0.99	0.95
BAEE	2.33	**2.81**	2.15	2.18	1.22

One may argue that the accuracy for BAEE and AEE is always lower than that of B-AUB and AUB. This fact can be expected because the uncertain area in BAEE and AEE is theoretically smaller. The strongest advantage of BAEE and AEE is that they allow a tradeoff between accuracy and precision, which is reflected by the area of uncertain regions. BAEE and AEE can be deemed effective if such a tradeoff is cost-efficient and affordable in practice. However, this advantage becomes less significant, especially when higher weights are put on the accuracy rather than precision. Nevertheless, when we purely consider the proportion of reduced area and decreased accuracy without weighting them, the tradeoff is worthy, and the proposed models are generally better than traditional ones.

5.4.3 Sensitivity analysis

To further explore the sensitivity of the proposed methods regarding different initial parameters, that is, the sampling rate and the magnitude of measurement error, a sensitivity analysis was conducted on Geolife, Indoor and Bus datasets. The results are plotted in Figure 5.9. On the one hand, as the sampling rate increases, AEE and BAEE become less efficient in reducing the uncertain region (first column in Figure 5.9). This finding is reasonable because a higher sampling rate indicates more

substantial uncertainty in the raw trajectory. Consequently, the movement pattern embodied in the uncertain trajectory may be unrealistic; thus a conservative p value, which corresponds to large error ellipse, tends to be obtained. On the other hand, AEE and BAEE models achieve a slightly lower accuracy, which is embodied by the raw trajectory points coverage, compared with AUB and B-AUB model. The slight difference (less than 2% in most cases) proves the effectiveness of AEE and BAEE in sacrificing a little accuracy to reduce a large proportion of uncertain region area (second columns in Figure 5.9). Finally, we compare the sensitivity of BAEE and B-AUB (regress to AEE and AUB when the measurement error is zero) in terms of different measurement errors. As the measurement error increases, the accuracy of BAEE and B-AUB increases, whereas that of BAEE is slightly lower that of B-AUB (third columns in Figure 5.9). In that sense, we encourage to consider measurement error in practice when constructing a trajectory uncertainty model, which is expected to produce realistic results.

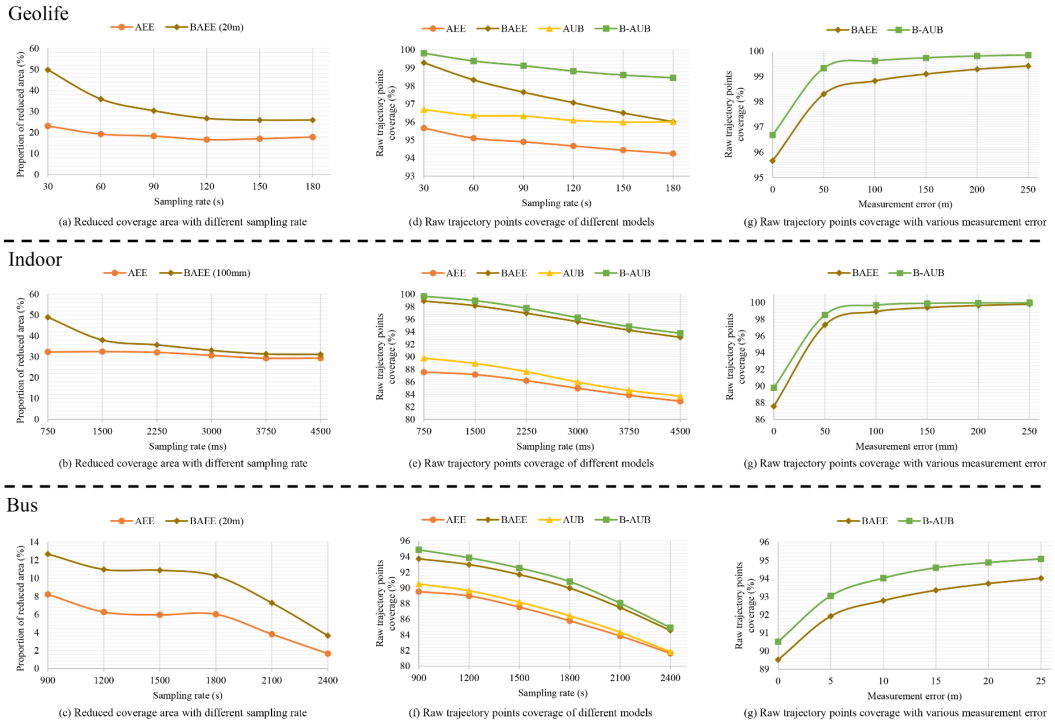
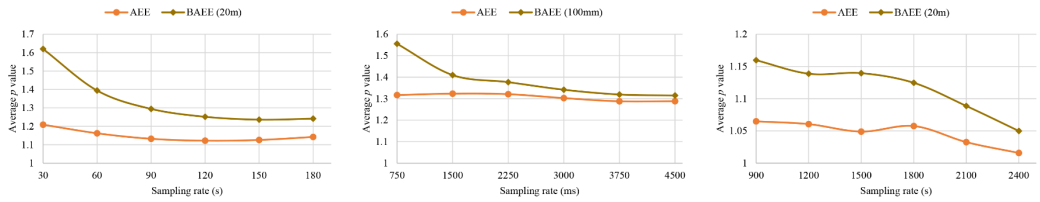


Figure 5.9 Sensitivity analysis for different models with various sampling rates and measurement error.

5.4.4 Effectiveness of optimisation process

The influences of the sampling rate on the optimisation process are initially discussed. A high sampling rate leaves a high uncertainty to the resampled trajectory. Thus, a broad uncertain region (i.e. smaller p value) is required to cover the actual movement points. The average optimal p values for Geolife, Indoor and Bus datasets with different sampling rates are plotted in Figure 5.10. For AEE and BAEE models, the average optimal p value slightly decreases with the increase of the sampling rate. This result is consistent with our expectation, proving the effectiveness of the optimisation process in adaptively providing smaller p for low sampled data.



(a) Geolife (b) Indoor (c) Bus

Figure 5.10 Effects of sampling rates on optimisation results.

The optimised results for Cabs A and B, which contain relatively homogeneous data (i.e. produced by ride-hailing service from the same company) from two cities with different road network structures, are compared to further validate the effectiveness of the optimisation process. Maps of the road network in the urban area of Xi'an and Chengdu are shown in Figure 5.11, of which the background maps are retrieved from Google. As an ancient capital of China, the road network of Xi'an preserves the boxy style, and most roads present a horizontal or vertical direction. By contrast, the road network in Chengdu is more heterogeneous, and the overall style is close to multiple rings centring at the same spot. Therefore, the optimal p values for Xi'an and Chengdu are expected to be different.

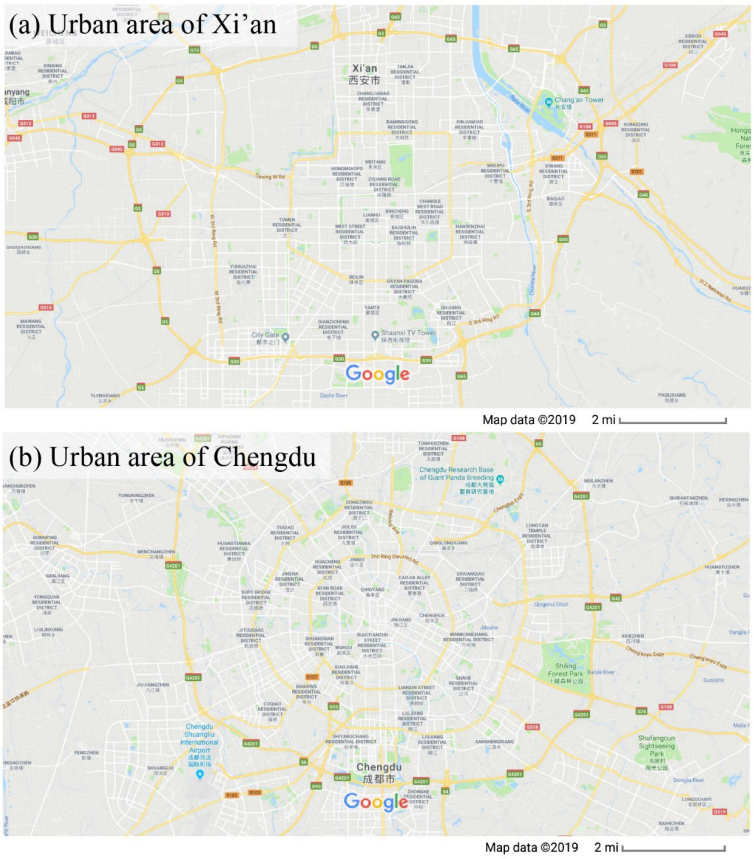


Figure 5.11 Maps of the urban area in Xi'an and Chengdu. Data are retrieved from Google Map.

The distributions of optimal p values for the two cab datasets with the same parameter setting in Section 5.4.2 (i.e. 60 s for resampling and 20 m for measurement error) are depicted in Figure 5.12. For AEE and BAEE models, optimal p values visually distribute differently for the two datasets. That is, more p values for Cab A (Xi'an) gathers close to 1, which corresponds to the Manhattan distance, whereas the distribution of p for Cab B (Chengdu) is flatter, which is consistent with our expectation. Finally, on the basis of the Mann–Whitney U test, which is a nonparametric test for independent samples, the p values of Cabs A and B are significantly different at a 95% confidence level as shown in Figure 5.12.

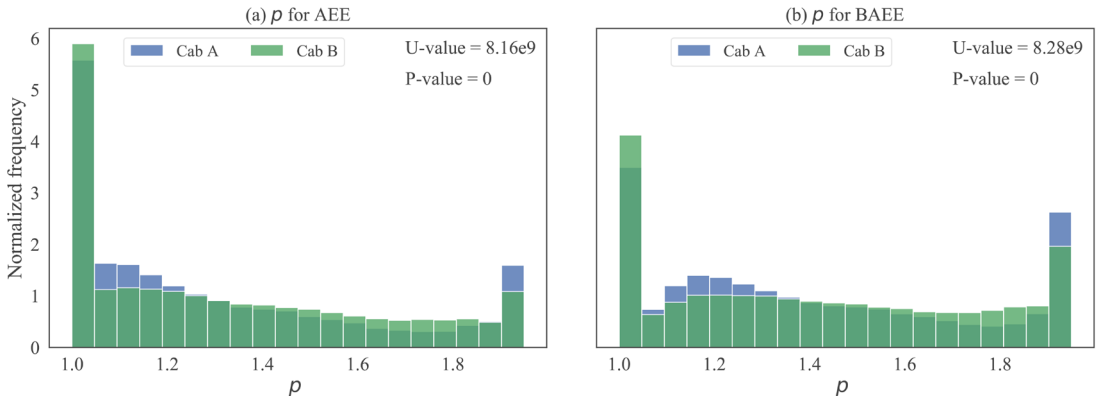


Figure 5.12 Distribution of optimal p values in Cabs A and B.

5.4.5 Case study on trajectory similarity analysis

We conduct a case study on trajectory similarity analysis based on the measure proposed in the original paper of AUB method, that is, the uncertain movement similarity (UMS), which measures the similarity of two trajectories by comparing their uncertainty ellipses. In this manner, we can exemplify the advantages of the proposed models in practical applications. The detail of UMS can be found in the literature (Furtado et al. 2018). The experimental process is designed similarly to the one used in the literature, which assumes that the trajectories between two given points of interest (POIs) should have higher similarity measurements than those between other POIs. The experimental process includes three steps.

1) Select all the trajectories from one specific POI to another labelled by the POI's name (e.g. L0–L1 if from L0 to L1), and form a trajectory set T. Given n pairs of POIs, n trajectory sets are obtained, which is expressed as $D = \{T_1, T_2, \ldots, T_n\}$.
2) For each trajectory, its similarity with the remaining trajectories in D is calculated using UMS with AUB, AEE and BAEE.
3) For the resultant similarity measurements of each trajectory in D, calculate the precision for different levels of recall by taking all the trajectories with the same label (e.g. L0–L1) as positive samples.
4) The average precision for each trajectory is recorded as the results.

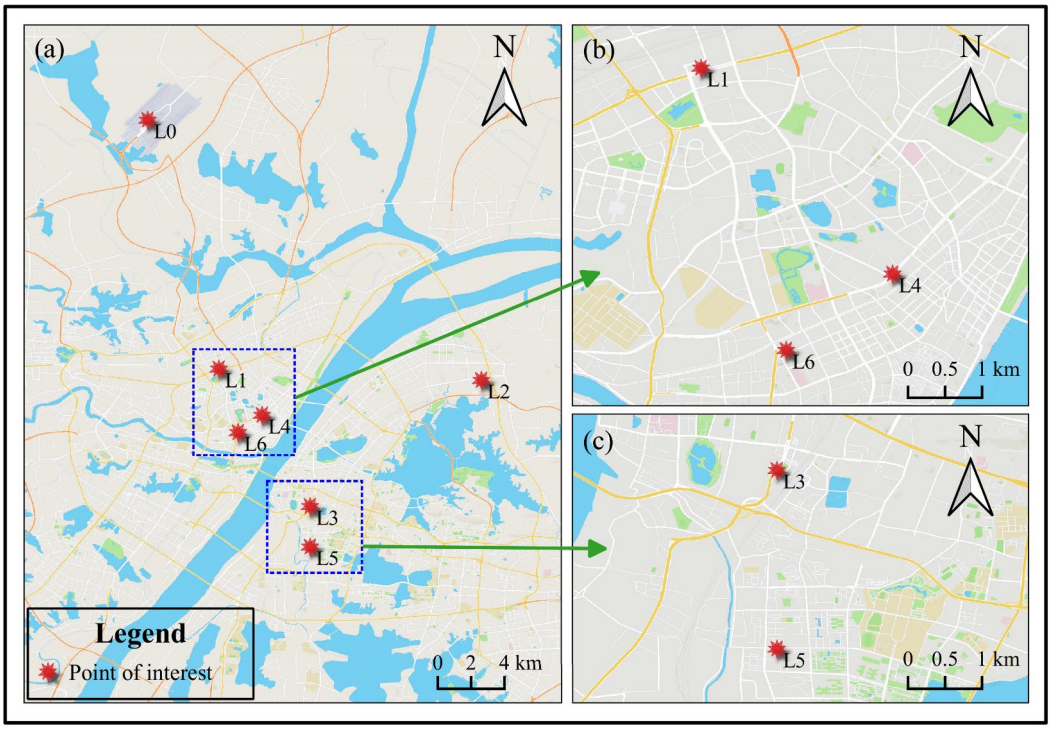


Figure 5.13 Spatial distribution of selected POIs. (b) and (c) are the enlarged maps that show the road structures in different regions.

On the basis of the original paper of UMS, overestimated ellipses tend to cause false-positive ellipse interaction, which reduces the overall precision of the similarity measurement (Furtado et al. 2018). Therefore, using the AEE and BAEE model instead of the AUB model in UMS is expected to improve the precision. Although the performance of UMS can be improved by other means, it is out of the scope of this study. We focus on the influence solely brought by different uncertainty models.

A taxi trajectory dataset collected in Wuhan, China was used in the case study. As shown in Figure 5.13, seven representative POIs (L0–L6) and fourteen routines with the highest number of trajectories amongst these POIs are selected for the similarity analysis. The details of these POIs and the statistics of the selected trajectories are listed in Table 5.4 and Table 5.5, respectively. Using the mean average precision (MAP) of UMS implemented by AUB as the baseline, the MAP improvement brought by AEE and BAEE is shown in Figure 5.13 using boxplots grouped by the trajectory sets. The measurement error in BAEE is set to 50 m.

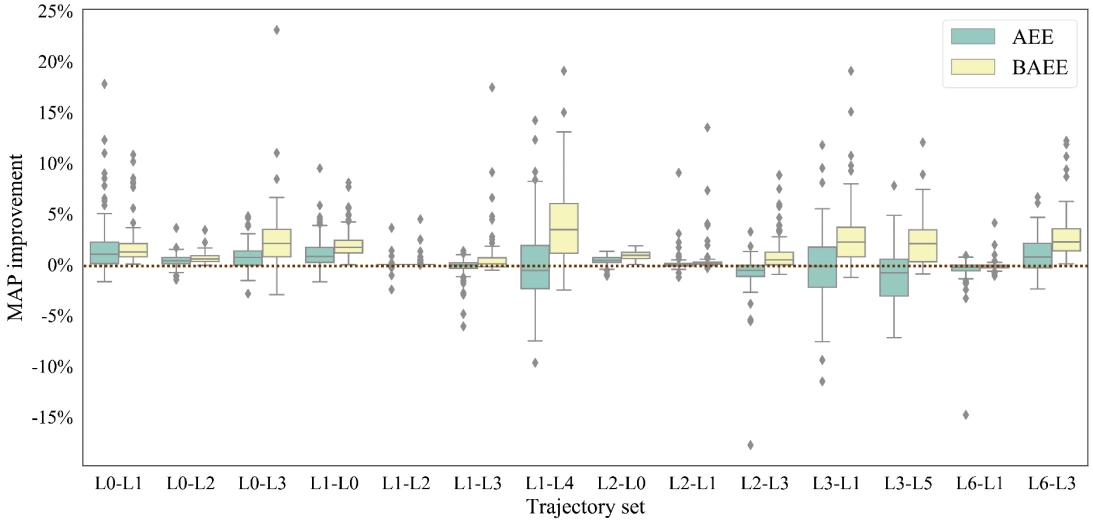


Figure 5.14 Boxplots of the MAP improvement achieved by AEE and BAEE for each trajectory set in UMS analysis. The improvement is measured by regarding the MAP achieved by AUB as the baseline.

As shown in Figure 5.13, AEE and BAEE increase the MAP on average for all trajectory sets. The overall improvement of MAP verifies our expectation that the AEE and BAEE models can alleviate the overestimation in AUB. However, a small part of trajectories (points below the dashed line) shows a decreased MAP using AEE or BAEE. We investigate such exceptions and conclude the reason as follows. Given the unstable sampling interval for each trajectory and measurement error, some raw trajectories may be a nearly straight line, especially for the short routines, such as L1–L4 and L3–L5, which results in a large Minkowski coefficient p and makes the uncertain region, especially for AEE model, extremely narrow to intersect with others. Therefore, the AEE model in such situations might have relatively unstable performance, as observed in Table 5.4 for trajectory sets L1–L4 and L3–L5. Conversely, BAEE performs considerably better by including the measurement error.

Table 5.4 Descriptions of selected POIs.

POI	Category	Description
L0	Airport	Tianhe International Airport
L1	Train station	Hankou Train Station
L2	Train station	Wuhan Train Station

L3	Train station	Wuchang Train Station
L4	Commercial centre	Jianghan Road
L5	Residential area	South Lake Community
L6	Commercial centre	Wuhan Square

Table 5.5 Statistics of selected trajectories.

ID	Label	From	To	Number of trajectories	Avg. points per trajectory	Avg. sampling distance	Avg. sampling time
1	L0–L1	L0	L1	170	26.89±5.24	884.96±143.88	66.91±5.26
2	L0–L2	L0	L2	84	36.80±3.46	1193.58±103.52	66.50±4.74
3	L0–L3	L0	L3	45	45.31±6.19	983.25±208.44	68.45±5.04
4	L1–L0	L1	L0	156	27.12±4.16	930.47±110.02	66.33±4.89
5	L1–L2	L1	L2	113	25.44±13.39	927.63±178.70	64.39±9.67
6	L1–L3	L1	L3	89	24.17±5.34	629.63±103.76	69.41±6.74
7	L1–L4	L1	L4	65	17.38±6.30	378.59±176.70	65.21±6.79
8	L2–L0	L2	L0	74	39.03±5.68	1159.62±136.26	66.98±4.83
9	L2–L1	L2	L1	112	25.22±4.43	893.25±133.34	66.49±4.46
10	L2–L3	L2	L3	103	30.75±7.14	649.55±171.02	66.42±4.75
11	L3–L1	L3	L1	60	27.00±8.40	606.78±147.66	68.07±6.15
12	L3–L5	L3	L5	57	16.39±9.28	291.91±101.83	61.70±10.14
13	L6–L1	L6	L1	161	13.58±4.60	453.67±140.39	64.71±9.23
14	L6–L3	L6	L3	40	20.05±12.57	578.70±150.49	68.45±14.73

We perform a paired T-test to further validate our observation from the statistical perspective. The results are listed in Table 5.6. BAEE achieves the best MAP, followed by AEE, and the improvement is significant at the 95% confidence interval. The test results prove the effectiveness of BAEE and AEE in this case study, and MAP improvement of BAEE against AUB is over 1.3% on average in the case shown in Table 5.6.

Table 5.6 Paired T-test of the MAP of different uncertainty models in the case study.

Tested pair	Paired differences					Sig (Two-tailed)
	Mean	SD	Std. error mean	95% confidence interval for the mean difference		
				Lower	Upper	
AEE–AUB	0.0024	0.0217	0.0004	0.0016	0.0033	0
BAEE–AUB	0.0133	0.0233	0.0005	0.0124	0.0141	0
BAEE–AEE	0.0111	0.0271	0.0005	0.0101	0.0122	0

5.5 Summary

This chapter introduced a novel trajectory uncertainty model. Three contributions were made in this chapter. 1) An AEE model was proposed by extending the beads model. The AEE model no longer requires the speed information but trains a hyperparameter for Minkowski distance metric based on underlying mobility patterns embodied in each trajectory. In this manner, the thickness of error ellipses can be determined adaptively. 2) A BAEE model, which highlights the measurement error by adding an uncertain range for each sampled point in the AEE model, was further developed. BAEE has no standard mathematical expression. Thus, we deduce an approximate error ellipse for its representation. 3) Experiments were performed on five real-life datasets to show the superiority of the proposed models in providing a more reliable boundary for the actual movement than the state-of-the-art approaches. A case study was also made on a taxi trajectory dataset for similarity analysis to exemplify the value of the proposed models in practical applications.

On the basis of the experiments and case study, the advantages of the proposed models can be concluded as follows. 1) AEE and BAEE models could significantly reduce the overestimated uncertain regions produced by state-of-the-art approaches and provide a more reliable uncertain boundary for modelling the actual movement. 2) The proposed optimisation process could efficiently capture the potential network constraint based on the raw trajectory without additional information. 3) AEE and BAEE models performed robustly with respect to different sampling rates, sizes of measurement error and moving patterns. 4) BAEE and AEE model could improve the precision of the state-of-the-art trajectory similarity analysis approach, whereas BAEE overcame the problem of extremely narrow ellipses generated due to straight-like trajectories and achieved better precision in the case study.

The AEE and BAEE model are theoretically more rational and realistic than the state-of-the-art methods, which promotes a bright prospect in related applications. One limitation for this study is each trajectory is associated with a specific optimal distance

metric, while the optimal result may be affected by the movement uncertainty as well. Based on our assumption that movement is restricted by a latent network, such optimal distance metric should be consistent within a local area. Therefore, our future work will focus on designing more sophisticated algorithms for mining the best distance metric for specific regions.

Chapter 6 Exploring the uncertainty in time-series big data and its application in human behaviour prediction

In addition to trajectory data, spatial time-series is another representative SBD type that is ubiquitous in modern human behaviour studies, such as mobility pattern mining and public transit usage forecasting. Due to the randomisation in human behaviour, time-series data may present different levels to be modelled and forecasted, which is also called predictability in literature. The predictability of time-series data can considerably affect the reliability of related modelling and forecasted results. In that sense, understanding the predictability in spatial time series can help build a sense of control for a project manager and contribute to reducing the risk of unexpected loss. This chapter takes the metro ridership data in Shenzhen, China, as an example, introduces a novel predictability evaluation method for spatial time series data and demonstrates the practical usage of the predictability measurements in enhancing the prediction performance of deep learning algorithms.

6.1 Overview of the predictability of time-series data

A time series consists of a series of data points collected in time order. Time-series data are nearly omnipresent in any scientific domain that involves temporal dimension. Through specific time-series analytics, rich underlying characteristics and regularity of the observed data can be extracted to support further applications. At present, time-series data have widely been used in temporal pattern mining and forecasting (Hamilton 1994, Chatfield 2000).

Predictability refers to the degree that a system's state can be correctly modelled or forecasted (Colwell 1974). Predictability of time-series data has long been recognised as a significant property and received specific concerns from many scientific communities, such as economics, ecology, epidemiology and earth science (Campbell and Shiller 1988, Stephens 1993, Prank et al. 1995, Scarpino and Petri 2019). Predictability is also a popular area of human behaviour research. A prominent

example can be seen in the work of Song et al. (2010), which introduced a real entropy index for individual movements and utilised Fano's inequality to measure the upper bound of predictability of mobile carriers. This work also led a series of studies on the quantification of predictability in human mobility (Lu et al. 2012, Marin et al. 2014, Smith et al. 2014, Yan et al. 2014).

Previous studies always take predictability as the inherent property of time-series data that are independent of the algorithms or models selected for the prediction task (Garland et al. 2014). For this reason, considerable efforts have been made to construct useful indicators for quantifying predictability. Entropy plays an essential role in existing time-series predictability indicators. Classical entropy indicators for time-series data, such as ApEn, sample SampEn, multiscale entropy and permutation entropy (Pincus 1991, Richman and Moorman 2000, Costa et al. 2002, Borowska 2015), have been applied in a broad range of studies, providing quantitative predictability measurements solely based on the intrinsic regularity in time-series data.

Into the era of big data, modelling and forecasting human behaviour using various time-series data, which benefit from long-term observations through continuous surveillance by portable GPS and smart devices, becomes pervasive. Certain entropy indicators specially designed for spatial time series are available. Xu et al. (2017) established an entropy-based method for estimating travel time predictability by implementing the multiscale entropy of travel time series into Fano's inequation. Chen et al. (2019) proposed a SampEn-DNN approach, in which the SampEn was adopted to represent the predictability information for Bulletin Broad System time-series forecasting. In addition to entropy-based indicators, other techniques, such as sample autocorrelation (Musolesi and Mascolo 2006), Poisson process (Ihler et al. 2007) and F1 score (Foell et al. 2015), have also been used for measuring predictability.

However, from the perspective of decomposition, the predictability of a time series primarily stems from the uncertainty of each decomposed component, as well as their interactions. As the decomposed components always present different

predictability (details will be given in Section 6.2.2), the above methods, which only measure the raw time series as a whole, may produce unrealistic measurements due to the underlying intervention amongst these components. On the other hand, some popular prediction models for time series, such as the seasonal autoregressive integrated moving average (SARIMA) model, also involve the decomposition technique (Vagropoulos et al. 2016). The sensitivity of these prediction models to different decomposed components can be inconsistent. Therefore, conducting standalone predictability analysis on the decomposed components is helpful to avoid the uncertainty brought by their intervention and consistent with the theoretical basis of these models to provide realistic predictability measurements.

6.2 Modelling the predictability of human mobility as a time series

6.2.1 ApEn and SampEn

Amongst the abovementioned entropy indexes, ApEn and SampEn are most representative, which have been broadly employed in related works (Richman and Moorman 2000, Šliupaitė et al. 2015, Cao and Lin 2018). Here, a brief introduction is given to help readers understand their mathematical basis.

ApEn is widely used to quantify the predictability of a time series. In ApEn, subseries of a specific length m makes up a new m-dimensional space. ApEn reflects the likelihood of the emergence of new patterns when the length increases from m to $m+1$ (Pincus 1991). Generally, ApEn is expressed as a nonnegative number, and a large ApEn indicates low predictability. The calculation steps of ApEn are given as follows:

- Step 1: Given a time series U of N data points, the N data values are expressed as $u(1), u(2), \dots, u(N)$.
- Step 2: Given a fixed integer m, a set of m-dimensional vectors can be constructed, $X(i) = [u(i), u(i+1), \dots, u(i+m-1)], 1 \le i \le N-m+1$.

- Step 3: Let $d[X(i), X(j)]$ be the maximum absolute difference between the respective scalar elements of vector $X(i)$ and $X(j)$. Mathematically, $d[X(i), X(j)] = \max|u(i+k) - u(j+k)|, k = 0,1, \dots, m-1$.
- Step 4: Given a similarity threshold r ($r > 0$), consider the number of vector pairs satisfying $d[X(i), X(j)] < r$ for each i and the total number of vectors. We define

$$C_i^m(r) = num\{d[X(i), X(j)] < r\}/(N-m+1), 1 \le i, j \le N-m+1. \quad (6.1)$$

- Step 5: Define the natural logarithmic mean of $C_i^m(r)$ as $C^m(r)$, which is expressed as

$$C^m(r) = \sum_{i=1}^{N-m+1} \log(C_i^m(r)) / (N-m+1). \quad (6.2)$$

- Step 6: Increase m to m+1. Repeat Steps 3–5 and obtain $C^{m+1}(r)$. ApEn is finally expressed as

$$\text{ApEn} = C^m(r) - C^{m+1}(r). \quad (6.3)$$

The performance of ApEn relies on the selection of the vector length m and the similarity threshold r. On the basis of the existing literature, m is commonly set to 2, and r is set to $0.1-0.25SD$ (*SD* refers to the standard deviation of the raw time series U; (Pincus 1991, Richman and Moorman 2000).

SampEn is another index developed based on ApEn. The calculation of SampEn is quite similar to ApEn with only slight modifications to avoid self-match in ApEn (Richman and Moorman 2000). The first three steps for computing SampEn are the same as those for ApEn, and the remaining steps are as follows.

- Step 1–3: Similar to Steps 1–3 in ApEn calculation
- Step 4: Given a similarity threshold r ($r > 0$), consider a number of different vectors satisfying $d[X(i), X(j)] < r, i \neq j$ for each i and the total number of vectors. Then, we define

$$B_i^m(r) = num\{d[X(i), X(j)] < r\}/(N-m+1), 1 \le i, j \le N-m, i \ne j. \quad (6.4)$$

- Step 5: Let $B^m(r)$ be the average of $B_i^m(r)$. Then, we have

$$B^m(r) = \sum_{i=1}^{N-m+1} B_i^m(r)/(N-m). \quad (6.5)$$

- Step 6: Increase m to m+1. Repeat Steps 3–5 and obtain $B^{m+1}(r)$. SampEn is defined as

$$\text{SampEn} = -\log[\frac{B^{m+1}(r)}{B^m(r)}]. \quad (6.6)$$

SampEn is also nonnegative, and a small value indicates high predictability. The selection of parameters for SampEn is similar to ApEn, that is, $m = 2$ and $r = 0.1-0.25\,SD$.

6.2.2 Introduction on time series decomposition

The decomposition technique is widely adopted in time series analysis in current studies (Verbesselt et al. 2010, Wang et al. 2015, Qiu et al. 2017). Decomposition technique can resolve a time series into multiple components based on their different predictability. Generally, a time series y_t can be decomposed into three parts (Hyndman and Athanasopoulos 2018):

- T_t: trend-cycle component, which refers to the general trend of the series (e.g. upwards and downwards)
- S_t: the seasonal component, which exhibits a periodical pattern (e.g. daily and weekly)
- R_t: residual component, which represents random noise after the extraction of other components

Hence, an additive model for the decomposition can be written as:

$$y_t = T_t + S_t + R_t. \quad (6.7)$$

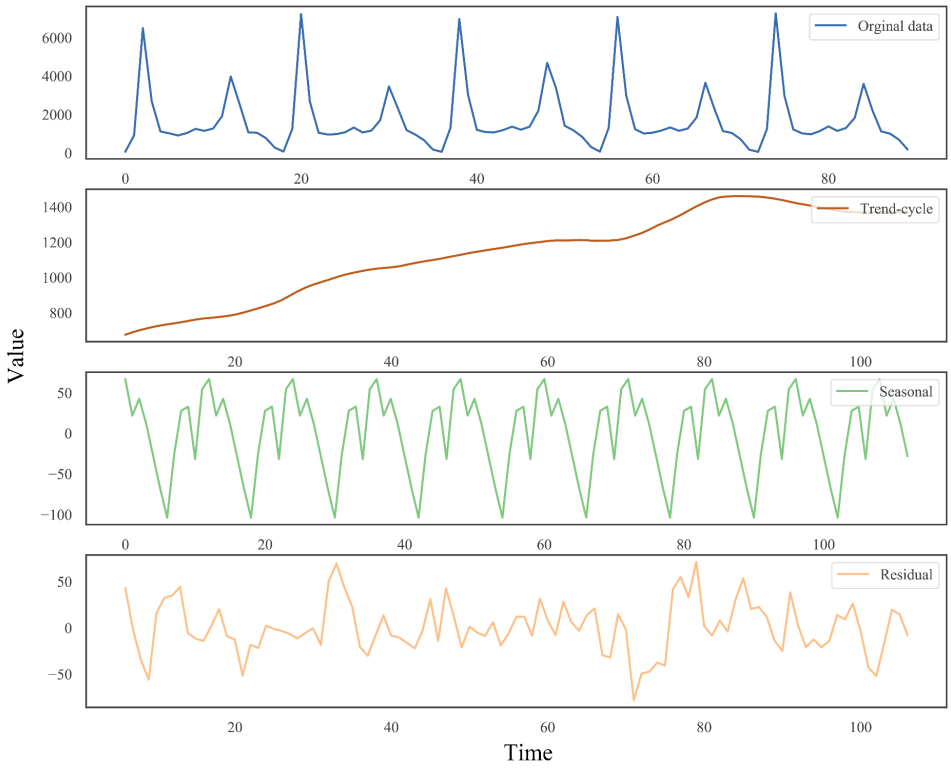


Figure 6.1 Example of the decomposition of time series.

An example is given in Figure 6.1 to illustrate the decomposition result based on the additive model. In comparison with the original data, the seasonal component exhibits a higher level of regularity, which makes it nearly predictable. By contrast, a modellable pattern can be observed in the trend-cycle component, whereas the residual component seems to be random.

6.2.3 Novel indicator for the overall predictability of a time series

The predictability of the raw time series (i.e. metro ridership) and the decomposed components under an additive decomposition model are considered when constructing the predictability indicator. Given a time series *TS*, we define a component set as $S_{TS} = \{s_{TS}^0, s_{TS}^1, s_{TS}^2, \dots, s_{TS}^m\}$, where s_{TS}^0 refers to the raw time-series data; and $s_{TS}^1, s_{TS}^2, \dots, s_{TS}^m$ represent m decomposed components. The raw data s_{TS}^0 is included because it can reflect the combined effects of the decomposed components, which may provide unique predictability information than other components. Given n such time series $\{TS_1, TS_2, \dots, TS_n\}$ and a primary entropy measure, an $n \times (m+1)$ entropy

matrix can be constructed as

$$M_{Ent} = \begin{bmatrix} H_0 & H_1 & \cdots & H_m \end{bmatrix} = \begin{bmatrix} h^0_{s_{TS_1}} & h^1_{s_{TS_1}} & \cdots & h^m_{s_{TS_1}} \\ h^0_{s_{TS_2}} & h^1_{s_{TS_2}} & \cdots & h^m_{s_{TS_2}} \\ \vdots & \vdots & \vdots & \vdots \\ h^0_{s_{TS_n}} & h^1_{s_{TS_n}} & \cdots & h^m_{s_{TS_n}} \end{bmatrix}, \tag{6.8}$$

where H_i refers to the entropy vector (EV), which consists of the entropy measurements of s^i for each time series.

This study assumes that the overall predictability of a time series TS is determined by the integration of entropy measurement of each item in the component set S_{TS}. On the basis of this assumption, taking the sum of entropy values as the final estimator is easy. However, given the potentially different significances and magnitudes of each component, simply adding all entropy values together may lead to biased results. For example, a component can have a high entropy value; however, it may have little influence on the overall predictability if it has a relatively low magnitude. To address this problem, we weigh the entropy values in terms of the ratio between the variation of each component and the raw data. Therefore, the overall predictability vector can be expressed as

$$H_{\text{overall}} = \sum_{i=0}^{m} H_i \times \frac{V_i}{V_0}, \tag{6.9}$$

where V_i is the variance vector $\{v^i_{TS_1}, v^i_{TS_2}, \dots, v^i_{TS_n}\}$ for ith item in the component set S.

In addition, according to our preliminary experiment, of which the details will be given in Section 6.2.4, the weighted EVs (WEVs) $H_i \times \frac{V_i}{V_0}$ can present various correlation strength regarding the actual predictability, which is measured by the prediction accuracy by specific models. Simply put, some WEVs can be unrepresentative for the actual predictability and make a negative contribution to the

final estimation. To reduce the uncertainty brought by unrepresentative WEVs during the estimation, we assume that the representative WEVs tend to have relatively similar estimations, whereas unrepresentative ones tend to produce different results. This assumption is reasonable because if two vectors are representative of the real predictability, then they should be strongly correlated. Thus, this study measures the similarity amongst WEVs, and the measurements will determine an additional weight for each WEV.

The similarities amongst WEVs are measured by the correlation distance (Székely and Rizzo 2014). The correlation distance is a robust distance metric for dimensionless variables. For vectors X and Y, the correlation distance is expressed as

$$d_{\mathrm{cov}}(X,Y) = 1 - \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} \quad , \tag{6.10}$$

where $\mathrm{cov}(X,Y)$ represents the covariances of X and Y; and σ_X and σ_Y refer to the standard deviation of X and Y, respectively. Thus, given a set of WEVs S_{WEV}, the vector weight w_{Vec^*} for a given item Vec^* in S_{WEV} is computed as

$$w_{WEV^*} = 1 - \frac{\sum_{WEV_i \in s_{WEV}} [d_{\mathrm{cov}}(WEV^*, WEV_i)]}{\sum_{WEV_i \in s_{WEV}} \sum_{WEV_j \in s_{WEV}, WEV_i \neq WEV_j} [d_{\mathrm{cov}}(WEV_i, WEV_j)]}. \tag{6.11}$$

On the basis of Equation (6.11), a relatively small weight will be generated if the given item is significantly different from others. By contrast, similar vectors will be taken as more reliable estimators, and larger weights are assigned. Therefore, on the basis of Equations (6.9) and (6.11), a weighted overall predictability vector $H_{\mathrm{w\_overall}}$ is defined as

$$H_{\mathrm{w\_overall}} = \sum_{i=0}^{m} w_i \times H_i \times \frac{V_i}{V_0}, \tag{6.12}$$

where w_i is obtained based on Equation (6.11). Theoretically, a large $H_{\mathrm{w\_overall}}$ corresponds to low predictability of the corresponding time-series data.

6.2.4 Approaching actual predictability

To validate the effectiveness of the proposed indicator, the real predictability is theoretically needed. However, we use the prediction accuracy of specific prediction models to approximate it due to the lack of efficient methods to directly obtain the actual predictability. Concerning the underlying uncertainty and potential bias in different prediction models, four prominent prediction models, namely, Holt-Winters model (Winters 1960), SARIMA with explanatory variable (SARIMAX) model (Brockwell et al. 1991), fbProphet (Taylor and Letham 2018) and long short-term memory (LSTM) neural network (Hochreiter and Schmidhuber 1997), of which the efficiency has been well proved in previous studies (Hyndman and Athanasopoulos 2018), are used in the experiment. Notably, although some advanced models exist, most of them improve the absolute prediction accuracy, whereas proposed indicators measure a likelihood that will not be significantly affected by slight changes of the absolute prediction accuracy.

The prediction accuracy of a time series can be measured by the root mean square error (RMSE). However, the mean ridership for each metro stations is not the same. Thus a normalised RMSE (NRMSE) is used instead, which can be expressed as follows:

$$\mathrm{NRMSE} = \frac{\mathrm{RMSE}}{\sigma}, \tag{6.13}$$

where σ denotes the standard deviation of the observed values in a prediction. Theoretically, the NRMSE measures the ratio between the unexplained variance by the forecasted results and the overall variance of the observed data (Otto 2019), which is conceptually consistent to the predictability, that is, the ability for data to be approximated and forecasted.

6.3 Experiments and analysis

6.3.1 Study area and data description

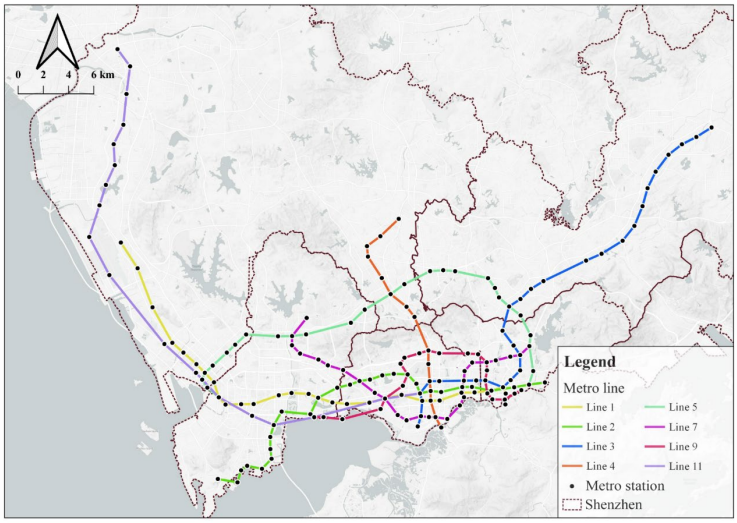


Figure 6.2 Metro network in Shenzhen, China.

As one of the most developed regions in China, Shenzhen plays a vital role in leading the practice of smart city. At present, eight metro lines, which consist of 166 metro stations that cover a large proportion of the urban regions, are running in Shenzhen, as shown in Figure 6.2. In this chapter, we adopt more than 70 million check-in/out records, spanning 25 days from January 2017 from the metro system in Shenzhen for the experiments. The check-in/out records are collected based on the transaction of the transportation smart cards. A typical record contains the station name, transaction time, card identification and some other auxiliary information. Table 6.1 lists the sample records.

Table 6.1 Samples for the metro check-in/out records. Trade type '21' stands for check-in, and '22' for check-out.

Card ID	Trade type	Station name	Time	Metro line ID
262020113	22	Cuizhu	2017-01-20 14:03:20	3
268012175	21	Lianhua West	2017-01-20 14:03:20	2
268008127	22	Shawei	2017-01-20 14:03:20	7
263036102	22	Shangtang	2017-01-20 14:03:20	4

The records are collected for 18 hours from 6 a.m. to midnight per day. The ridership data are summarised as the sum of check-in/out records during a specific time interval for each station. To explore the effect of temporal scale on the predictability of ridership data, multiple time intervals, including 20, 30 and 60 min, are applied to the raw records. On the basis of previous studies, significantly different patterns exist in the metro ridership between weekdays and weekends. Only the ridership data of 25 weekdays are studied to make the data homogenous for further analysis. Therefore, given the three intervals (i.e. 20, 30 and 60 min) and different trade types (i.e. check-in and -out), six datasets are obtained for further predictability analysis in this study and tagged by the trade type and interval length for easy referencing. For example, the check-in dataset with a time interval of 20 min is termed as In-20. The details of these datasets are listed in Table 6.2.

Table 6.2 Description of the six datasets in this study.

Tag	Trade type	Interval length (min)	Sample number
In-20	Check-in	20	1,350
In-30	Check-in	30	900
In-60	Check-in	60	450
Out-20	Check-out	20	1,350
Out-30	Check-out	30	900
Out-60	Check-out	60	450

6.3.2 Performance of the proposed method

Table 6.3 Main parameters for the four selected models.

Model	Meaning of parameters	Type
Holt-Winters	m: seasonal period	Integer
SARIMAX	p: trend autoregression order	Integer
	d: trend difference order	Integer
	q: trend moving average order	Integer
	P: seasonal autoregression order	Integer
	D: seasonal autoregression order	Integer
	Q: seasonal autoregression order	Integer
	m: seasonal period	Integer

fbProphet	ds: daily seasonality	Bool
	sm: seasonality model	'additive'; 'multiplicative'
LSTM	n: number of hidden units	Integer

To initialise the prediction, the former 80% data are used for training, and the remaining 20% data for validation. The optimal parameters for each model are carefully tuned before the prediction to reduce the effect of model uncertainty in our experiment. The main parameters, as well as their practical meanings, are listed in Table 6.3. Details about these parameters can be found in the literature (Vagropoulos et al. 2016, Donges 2018, Taylor and Letham 2018). The parameters finally used for each dataset are summarised in Table 6.4.

Table 6.4 Parameter settings in the experiment.

Dataset	**Holt-Winters (m)**	**SARIMAX: (p, d, q), (P, D, Q, m)**	**fbProphet (ds, sm)**	**LSTM (n)**
In-20	54	(2, 0, 2), (1, 2, 2, 54)	(True, 'multiplicative')	1000
In-30	36	(1, 0, 2), (0, 2, 2, 36)	(True, 'multiplicative')	1000
In-60	18	(0, 0, 2), (0, 2, 2, 18)	(True, 'multiplicative')	1000
Out-20	54	(2, 0, 2), (1, 2, 2, 54)	(True, 'multiplicative')	1000
Out-30	36	(2, 0, 2), (1, 2, 2, 36)	(True, 'multiplicative')	1000
Out-60	18	(1, 0, 2), (0, 2, 2, 18)	(True, 'multiplicative')	1000

The relationship between the NRMSE and the entropy estimators is further explored. For easy reference, we use H_y, H_T, H_S and H_R to represent the EVs corresponding to the raw data, trend-cycle component, seasonal component and residual component, respectively. Particularly, due to the high predictability of seasonal component, as discussed before, H_S is excluded during the estimation to avoid its interference in calculating the vector weights. Notably, each EV is implemented and termed based on the entropy index used in the experiment. For instance, the EV implemented based on ApEn is expressed as an ApEn-based vector.

The correlation coefficients between NRMSE and H_y, H_T, H_R, H_{overall} and $H_{\text{w\_overall}}$ are shown in Figure 6.3. In most cases, ApEn-based vectors have a significantly higher correlation coefficient than SampEn-based ones, which indicates the generally better performance of ApEn in measuring the predictability of the

experimental data. Furthermore, the three EVs (i.e. H_y, H_T and H_R) are alternatively to be of the highest correlation in specific cases. For example, in Cases A, B and C marked in Figure 6.3, H_T, H_R and H_y achieves the largest coefficient, respectively. Therefore, singly using H_T, H_R and H_y can result in unstable predictability estimation. By contrast, H_{overall} and $H_{\text{w\_overall}}$ are more robust because the corresponding correlation coefficients remain at a relatively high level in most cases. In addition, the coefficients corresponding to $H_{\text{w\_overall}}$ are generally higher than those of H_{overall}, implying the effectiveness of the weighting process described in Section 6.2.3. Exceptions exist, such as Case D, where H_T shows a larger correlation coefficient than the others. However, given the existence of model uncertainty and the robust performance of $H_{\text{w\_overall}}$, such exceptions are acceptable, which cannot deny the general effectiveness of $H_{\text{w\_overall}}$.

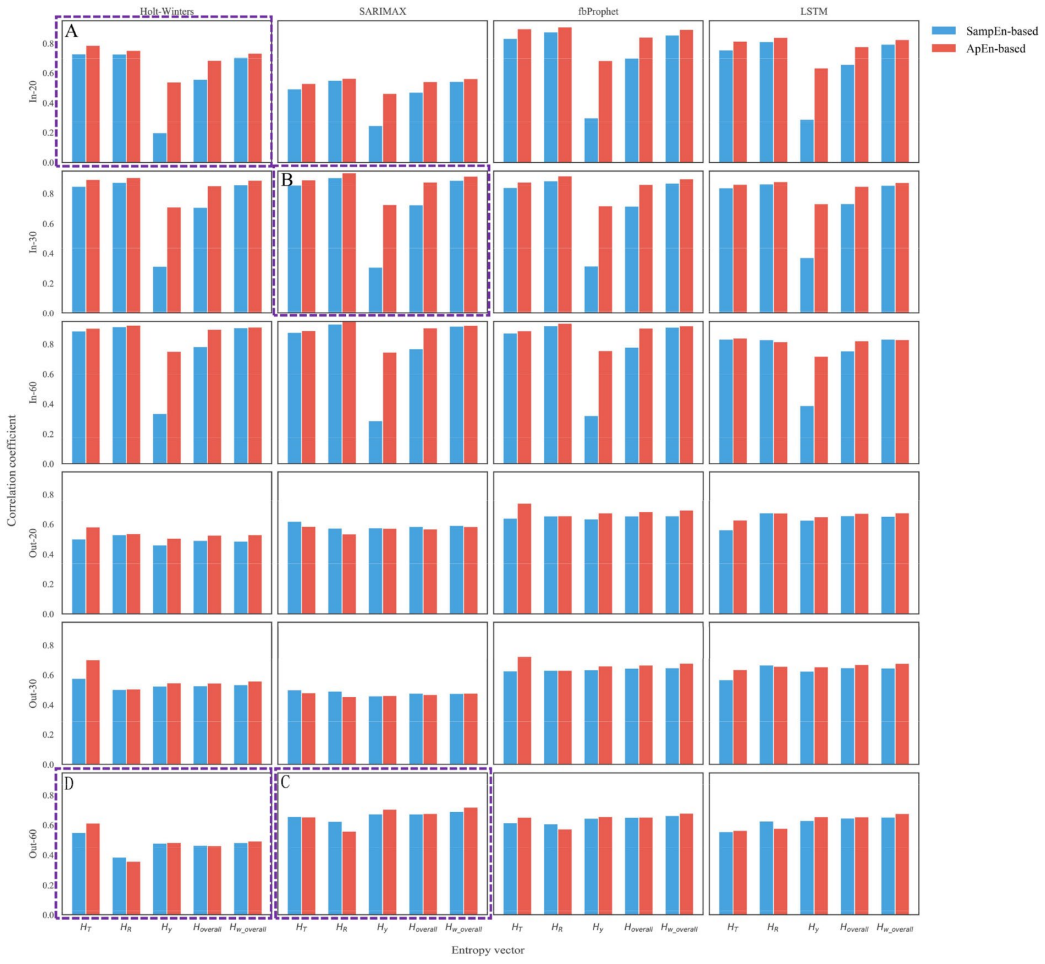


Figure 6.3 Correlation between EVs and NRMSE of different prediction models.

To further validate the effectiveness of the proposed indicator, the average correlation coefficient in terms of all those five ApEn-based vectors are taken as the baseline, and the ratios between the estimated correlation of each EV and this baseline are used to reflect the relative improvement on the correlation measurement.

Figure 6.4 shows the boxplots of the comparison results. Except for the Holt-Winters model, the high mean value and narrow box indicate the better performance of $H_{\mathrm{w\_overall}}$ than others. The exceptional case of Holt-Winters model, which is a primary prediction model, can be caused by its potentially higher model uncertainty. However, although the vector H_T exhibits the highest ratio in the case of Holt-Winters model, the large box indicates a strong fluctuance, implying its unstable performance in the case study. By contrast, the performance of the vector $H_{\mathrm{w\_overall}}$ remains stable and seems to be the second-best one amongst others in Holt-Winters case.

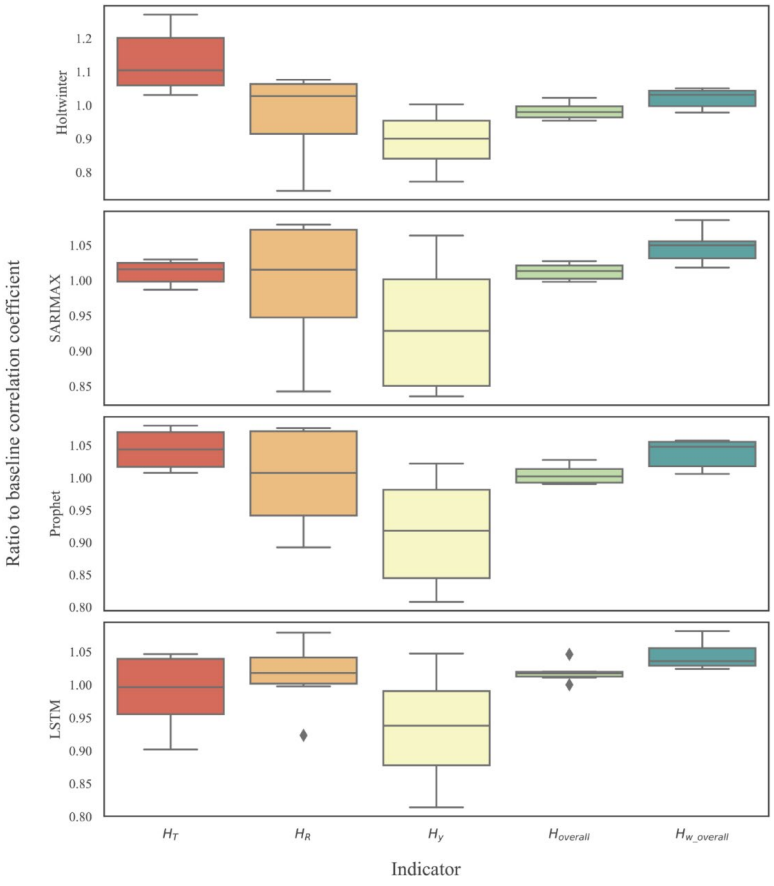


Figure 6.4 Boxplots for the ApEn-based vector on different prediction models.

As discussed in Section 6.2.1, the value of ApEn relies on parameters m and r. To investigate the effects of these parameters on $H_{\text{w\_overall}}$, m from 2 to 5 and r from 0.01 to 0.25 are applied to the In-30 dataset, and the corresponding correlation coefficients are plotted in Figure 6.5. Notably, for different values of m, the correlation coefficient keeps increasing with r before reaching a stable level. Furthermore, the value of m seems to have little influence on the results as the range of correlation coefficients for each prediction model is nearly the same with different values of m. Therefore, given the excess computation cost caused by larger m, $m = 2, r = 0.2$ is taken as a relatively good choice for the case in this experiment.

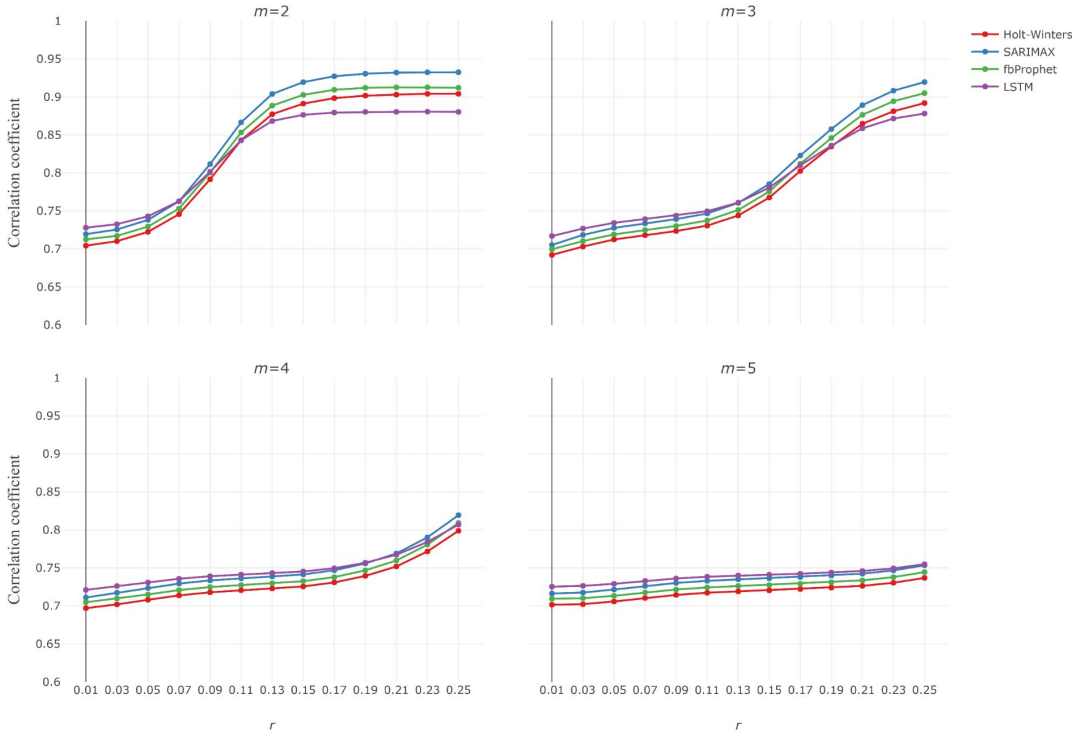


Figure 6.5 Effects of *m* and *r* on the performance of $H_{\text{w\_overall}}$ implemented by ApEn.

6.3.3 Result analysis

For convenience, the ApEn-based $H_{\text{w\_overall}}$ is expressed as ensemble ApEn (E-ApEn) for the rest of this discussion. The violin graphs of E-ApEn measurements in different datasets and the correlation matrix are plotted in Figure 6.6. For the check-in/out datasets, the range of E-ApEn narrows as the time interval increases. This result

is reasonable as the ridership series is abstracted with a large time interval. Thus, irregular fluctuance is discarded, which increases the predictability of less forecastable ridership series. In addition, the high values in the correlation matrix indicate that the estimations of E-ApEn are generally consistent at different temporal scales, that is, a large E-ApEn for a ridership series remains relatively large when the time interval changes.

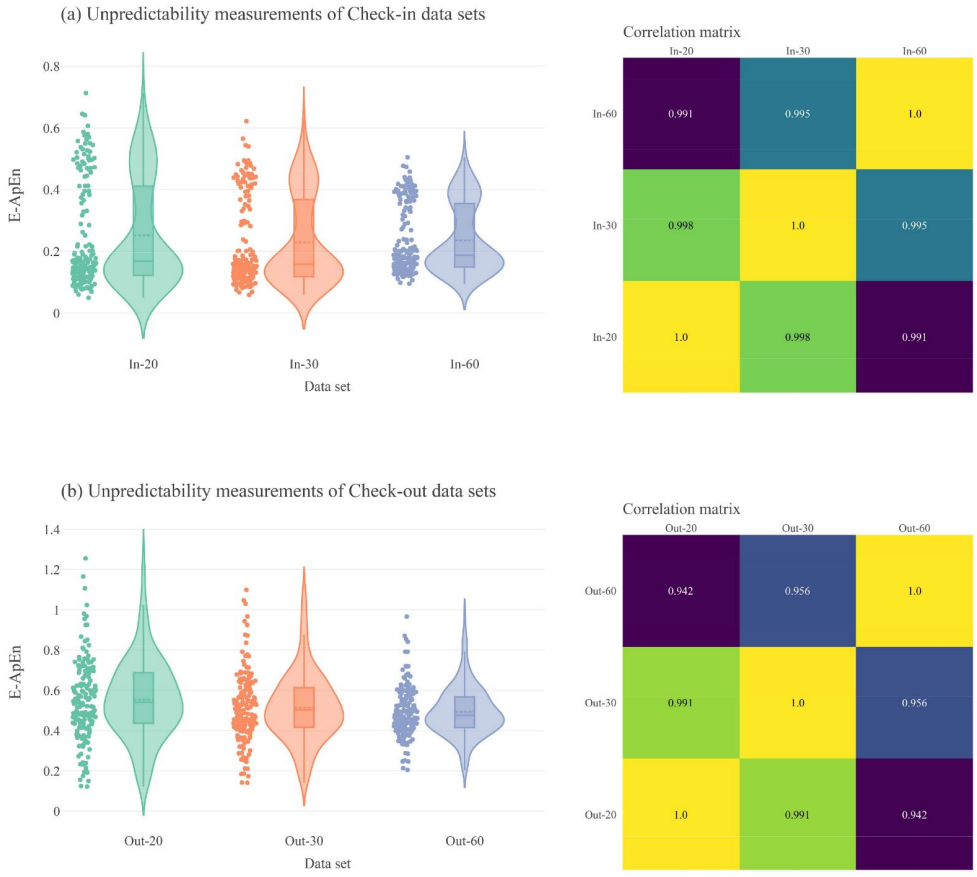


Figure 6.6 Distribution of E-ApEn measurements for different time intervals.

The E-ApEn measurements for the six datasets are mapped in Figure 6.7. Metro stations in some specific regions (e.g. the regions in the red boxes in Figure 6.7(a)) tend to have similar measurements. Intuitively, this local homogeneity is reasonable because adjacent or nearby metro stations may share similar ridership patterns because the daily activities near those stations can be quite related.

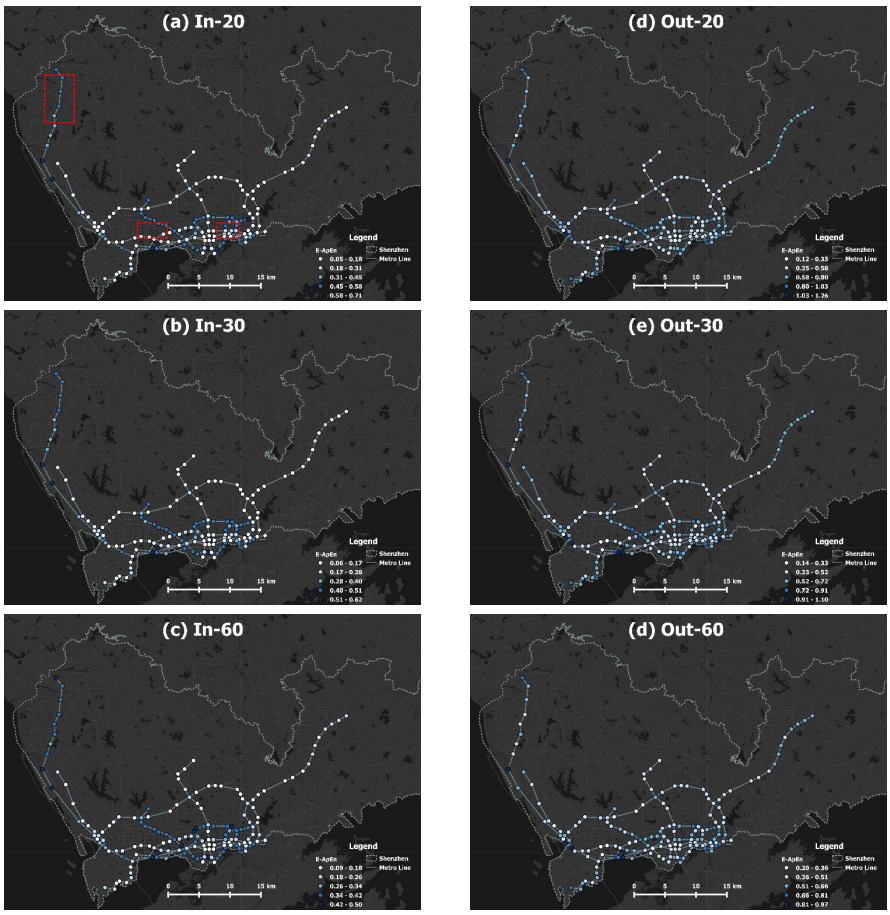


Figure 6.7 Maps of E-ApEn measurements for different datasets.

6.3.4 Application of predictability in deep learning

In comparison with traditional prediction models, deep learning models have shown advantages in time-series prediction in literature and real projects. Loss function plays an essential role in the training phase of deep learning models. The mean square error (MSE) is commonly used as the loss function in the deep learning model. For ridership prediction, let Y and $\hat{Y}$ respectively denote the predicted and actual ridership vector, which has a length of the number of metro stations. Then, the traditional MSE can be written as

$$\text{MSE}=\frac{1}{N}\sum_{i}^{N}(Y_i-\hat{Y}_i)^2, \tag{6.14}$$

where N is the size of validation data.

On the basis of the theory of weighted regression (Cleveland 1979), observations

may present different reliability; thus, applying appropriate weights on them can help improve the performance of regression. Similarly, in MSE loss function, the pre-estimated uncertainty of ridership data can facilitate the exploration of appropriate weights for better loss function. Inferring the uncertainty of ridership data of metro stations, which is conceptually close to predictability, is possible based on the proposed predictability indicator. In practice, the ridership with high predictability should have relatively large weight, whereas low predictability should be associated with a small weight. In that sense, we establish an uncertainty-based loss function.

Let $H = \{h_1, h_2, \dots, h_n\}$ denote the E-ApEn vector of n metro stations. The ridership uncertainty of metro station m can be calculated as the difference between h_m and maximum element in H represented by $\max(H)$. Then, to obtain standardised weights, the weight for metro station m is defined as

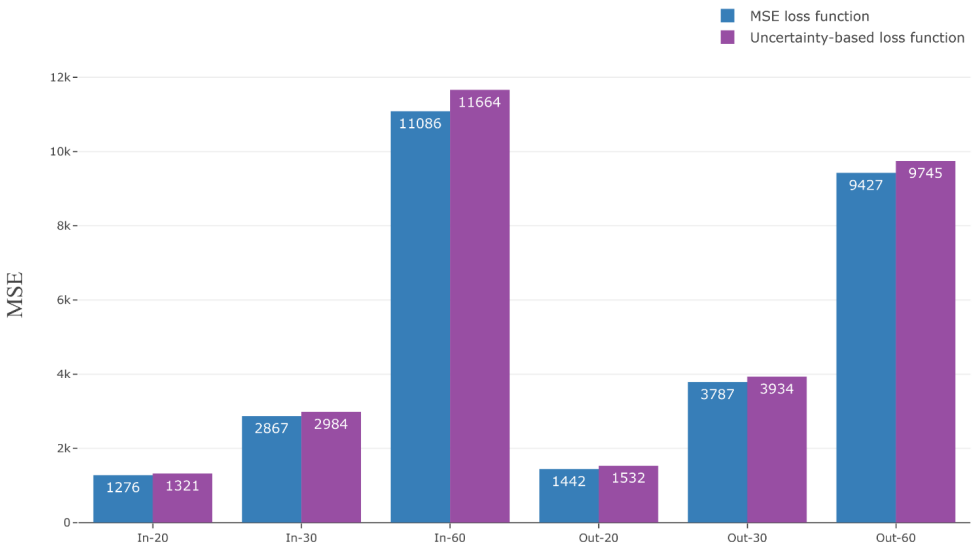
$$w_m = \frac{\max(H) - h_m}{\sum_{i}^{n} (\max(H) - h_i)}. \tag{6.15}$$

Let $W = \{w_1, w_2, \dots, w_n\}$ denote the weight vector, which comprises the weights of all metro stations. Given the dynamic training process, directly adding constant weights at the beginning of training may introduce bias to the rest of the training process. The ideal situation is to add weights to the loss function when the training process is about to converge. However, this process is unpredictable, such that constant weights are not suitable in such situations. Therefore, a dynamic exponential term $\lambda \times \text{epoch}$ is added to control the speed of W involved in the loss calculation, where λ is a constant that can be tuned. Therefore, the uncertainty-based loss function is defined as

$$\text{U-loss} = \frac{1}{N} \sum_{i}^{N} W^{\lambda \times \text{epoch}} (Y_i - \hat{Y}_i)^2. \tag{6.16}$$

The effectiveness of the uncertainty-based loss function is validated by applying

to LSTM and testing on the six datasets. The key parameters in this experiment (e.g. batch size number of recurrent units, epochs, dropout and the constant λ) are all carefully tuned to achieve the best prediction accuracy for each dataset, and early stopping criteria are used. The prediction accuracy improvements brought by uncertainty-based loss function in terms of MSE are shown in Figure 6.8. The uncertainty-based loss function has brought different degrees of improvement on prediction accuracy, which ranges from 3.26% to 5.88% on the experimental datasets, and the average improvement is 4.19%. The significant improvements reflect the effectiveness of the proposed uncertainty-based loss function. Many advanced deep learning models exist in current studies (Yao *et al.* 2018, Ma *et al.* 2019). Designing a better prediction model is out of the scope of our study. Thus, the uncertainty-based loss function is only tested on LSTM to validate its effectiveness, although it is theoretically applicable to the loss function in any other deep learning models.



Data set	In20	In30	In60	Out20	Out30	Out60
MSE improvement	3.40%	3.91%	4.95%	5.88%	3.74%	3.26%
Average improvement	4.19%					

Figure 6.8 Prediction accuracy of MSE and uncertainty-based loss functions.

6.4 Summary

This chapter demonstrated a new method for measuring the predictability of spatial

time-series data based on the decomposition technique and the traditional ApEn. Experiments were performed on a real-life ridership dataset from the metro system in Shenzhen to validate the effectivity of the proposed method. Four commonly used prediction models were applied to the same dataset to approximate the actual predictability. The experimental results showed that in comparison with classical indicators, the predictability measured by the proposed method exhibited a higher and stable correlation to the actual predictability, which proved its effectivity. Application of the predictability measurement was demonstrated by constructing an uncertainty-based loss function for improving the prediction performance of LSTM. In that loss function, dynamic weights were applied to the traditional MSE, where the selection of weights relied on the corresponding predictability measurements. The experiments showed that a significant improvement in prediction accuracy was achieved by the uncertainty-based loss function, which proved its effectiveness and the practical usage of the proposed evaluation method.

This study fills the research gap between the requirement of metro ridership predictability estimation and the lack of effective indicators. The well-performance of the proposed indicator in the case of Shenzhen proves its feasibility and theoretical correctness, promoting a prospect in its extension to other kinds of data such as bus ridership.

One limitation of this study is that only the well-known SampEn and ApEn were tested and compared, while some derivatives of these two indicators, as well as some other entropy indexes, such as real entropy (Song *et al.* 2010), multiscale entropy (Xu *et al.* 2017), should be investigated in follow-up studies. In addition, even some intuitive inferences were given to rationalize the effects of some factors on affecting the predictability of passengers' behavior, while the causality of some factors remained unclear, which requires more comprehensive and incisive work in the future.

Chapter 7 Conclusions and recommendations

7.1 Summary of this thesis

SBD has offered a new approach for comprehensively understanding the objective world, although SBD uncertainty is always amongst the most significant issues in related scientific research. On the one hand, traditional uncertainty analytics, represented by the quality assurance methods for geospatial products, are no longer suitable for voluminous, large-scale SBD. On the other hand, a growing demand exists for developing sound analytics for the uncertainty in newly emerging SBD types, such as human activity big data.

This thesis has responded to the current demands in developing better QAC methods, especially for large-scale spatial vector data, and sound uncertainty analytics for representative human activity data. As a result, efforts have been made as follows.

- A comprehensive review was conducted on the development of spatial uncertainty principle. An introduction was made on the current solutions and challenges in handling SBD uncertainty. Further discussions were made to emphasise the potential of data mining techniques in SBD uncertainty analysis and further guide studies in this thesis.
- A reference-free method was proposed to locate potential quality issues in multilayer vector data based on spatial relationship complexity. The linkage between complicated spatial relationship and quality issues was initially discussed. A contribution function was designed for measuring the complexity from the features in a single layer. Then, a new index revised from ENT was proposed to integrate the complexity contributions from multiple layers. On the basis of experiments on real-life and simulated data, the proposed method could provide a more realistic complexity measurement of the spatial relationship than the traditional approach.
- A reference-free method was proposed for detecting classification errors in

LULC data. In this method, the land patches were initially segmented by the MRS technique with an optimal scale, which was derived from the data characteristics, to reduce the uncertainty caused by patches aggregation. Then, the spectral and textural features were extracted from the resultant segments, and an adaptive clustering process was conducted to identify the outlying segments. Finally, on the basis of the statistics of the outlying segments and clustering results, an entropy-based index was constructed to evaluate the likelihood of a patch to be misclassified. Experiments were performed on two real-life datasets. Through comparison with state-of-the-art methods, the proposed method showed better performance on the experimental data in terms of higher TPR and PPV indexes.

- A novel approach was proposed to investigate the uncertainty brought by MMU in land classification data. On the basis of an assumption on the skew distribution of land patch size, multiple transformation functions were designed and iteratively applied to the raw data for normality correction. The transformed result, which corresponds to the best goodness of fit to general normal distribution, was selected to recover the omission error. Then, the commission error was estimated by the conversion probabilities amongst land classes, which were calculated based on the adjacency relationship amongst land patches. Finally, a confusion matrix could be constructed to evaluate the classification accuracy. Through the experiments on real-life and simulated datasets, the proposed reference-free method was proven to be feasible and effective.
- A novel approach was developed to address the movement uncertainty in trajectory data. Two uncertainty models were proposed, adopting elliptical regions with adaptive sizes to represent movement uncertainty. In the first model, the Minkowski distance metric was used to determine the size of error ellipse, and an optimisation process was designed to determine the optimal

Minkowski coefficient based on the characteristics mined from the raw trajectory. The second model extended the first model and further considered the measurement error during the calculation of the uncertain region. A standard formula for this model I unavailable. Thus, an approximated ellipse was deduced and validated through simulated data. The experiments demonstrated the superiority of the proposed models in narrowing the traditional error ellipse compared with the state-of-the-art method and exemplified their practical values in trajectory similarity analysis.

- A novel method was proposed to measure the predictability of spatial time-series data based on entropy and decomposition technique. Experiments were conducted on a 25-day metro ridership dataset from Shenzhen. To obtain the real predictability for comparison, four commonly used prediction models were applied to the same dataset, and the real predictability was represented by their normalised prediction accuracy (i.e. NRMSE). Through the correlation analysis, the proposed indicator was proved to provide more accurate and stable predictability estimations than traditional methods. In addition, the predictability measurements were implemented into an uncertainty-based loss function and applied to an LSTM model for ridership forecasting. The results showed that the novel loss function could bring an average improvement of 4.19% on prediction accuracy.

7.2 Limitations

Despite the current progress, this study still has certain limitations.

From the perspective of practical use, although some reference-free methods have been proposed, the manual work (e.g. visual inspection) and support materials (e.g. prior knowledge about the data), if possible, are still needed to validate the outputs of these reference-free methods. For example, in the method proposed in Chapter 4, the fundamental assumption may fail for some specific land classes due to

many factors, such as low data quality. To avoid this problem, a visual inspection of the data histogram should be initially conducted, and those land classes that do not satisfy a modellable distribution should be excluded.

This study also three limitations in terms of its methodology. On the one hand, in the reference-free method for classification error detection, only the spectral and textual features were adopted. As the features will be enriched with more image bands, adopting more features and their byproducts, such as NDVI, may help build a more distinguishable feature space to identify the outliers. On the other hand, regarding the predictability evaluation of spatial time series, only SampEn and ApEn were tested and compared, whereas other entropy indexes, such as permutation and multiscale entropies, should be investigated in follow-up studies.

From the theoretical perspective, although this thesis provides feasible solutions for certain fundamental uncertainty issues in SBD, more efforts are required to construct a comprehensive system for SBD uncertainty. Furthermore, this thesis exerted the most efforts on the uncertainty in data collection and processing, whereas less work was performed on the uncertainty in SBD analysis. SBD analysis undoubtedly involves various data types and a vast number of algorithms and applications. Thus, studying its uncertainty is promising and will significantly increase the applied value of SBD.

7.3 Recommendations

Recommendations for future works are given from three points of view.

From a theoretical perspective, a systematic uncertainty theory for SBD needs to be further developed. On the one hand, new uncertainty metrics are required for the uncertainty evaluation of various SBD types. On the other hand, more efforts should be given to the rationality and causality of human behaviour, which will promote the analysis and interpretation of the uncertainty in human activities.

From the methodological point of view. Artificial intelligence (AI) has great

potential for further developing reference-free methods for the QAC of earth observation big data. Through the continuous self-learning of AI on large-scale earth observation samples, sophisticated models can be generated to promote accurate and robust reference-free methods.

From the practical point of view, designing uncertainty-based approaches is always the dominant research direction, which provides a platform for SBD uncertainty theory to achieve its practical values. In future work, more efforts should be given on how to employ the uncertainty information at hand to improve current methods.

References

Ali, A.L., Falomir, Z., Schmid, F., and Freksa, C., 2017. Rule-guided human classification of Volunteered Geographic Information. *ISPRS Journal of Photogrammetry and Remote Sensing*, 127, 3–15.

Anselin, L., Syabri, I., and Kho, Y., 2006. GeoDa: an introduction to spatial data analysis. *Geographical analysis*, 38 (1), 5–22.

Antoniou, V. and Skopeliti, A., 2015. Measures and indicators of VGI quality: An overview. *ISPRS annals of the photogrammetry, remote sensing and spatial information sciences*, 2, 345.

Arrow, K.J., 1978. Uncertainty and the welfare economics of medical care. *In*: *Uncertainty in Economics*. Elsevier, 345–375.

Baatz, M. and Schäpe, M., 2000. Multiresolution segmentation-An optimization approach for high quality multi-scale image segmentation. *Angewandte Geographische Informations-Verarbeitung XII. Wichmann Verlag, Karlsruhe*, 12–23.

Baczkowski, A.J. and Clark, I., 1981. Practical Geostatistics. *Journal of the Royal Statistical Society. Series A (General)*, 144 (4), 537.

Baraldi, A., Bruzzone, L., Member, S., and Blonda, P., 2005. Quality Assessment of Classi cation and Cluster Maps Without Ground Truth Knowledge. *IEEE Transactions on Geoscience and Remote Sensing*, 43 (4), 857–873.

Barron, C., Neis, P., and Zipf, A., 2014. A comprehensive framework for intrinsic OpenStreetMap quality analysis. *Transactions in GIS*, 18 (6), 877–895.

Bartier, P.M. and Keller, C.P., 1996. Multivariate interpolation to incorporate thematic surface data using inverse distance weighting (IDW). *Computers and Geosciences*, 22 (7), 795–799.

Batty, M., 2013. Big data, smart cities and city planning. *Dialogues in Human Geography*, 3 (3), 274–279.

Bjørke, J.T., 1996. Framework for entropy-based map evaluation. *Cartography and Geographic Information Science*, 23 (2), 78–95.

Bonchi, F., Lakshmanan, L.V.S., and Wang, H. (Wendy), 2011. Trajectory Anonymity in Publishing Personal Mobility Data. *SIGKDD Explor. Newsl.*, 13 (1), 30–42.

Bora, M., Jyoti, D., Gupta, D., and Kumar, A., 2014. Effect of different distance measures on the performance of K-means algorithm: an experimental study in Matlab. *arXiv preprint arXiv:1405.7471*.

Borowska, M., 2015. Entropy-Based Algorithms in the Analysis of Biomedical Signals. *Studies in Logic, Grammar and Rhetoric*, 43 (1), 21–32.

Brockwell, P.J. and Davis, R.A., 2016. *Introduction to time series and forecasting*. Springer.

Brockwell, P.J., Davis, R.A., and Fienberg, S.E., 1991. *Time Series: Theory and Methods: Theory and Methods*. Springer Science & Business Media.

Brščić, D., Kanda, T., Ikeda, T., and Miyashita, T., 2013. Person tracking in large public spaces using 3-D range sensors. *IEEE Transactions on Human-Machine Systems*, 43 (6), 522–534.

Bruzzone, L., Cossu, R., and Vernazza, G., 2004. Detection of land-cover transitions by combining multidate classifiers. *Pattern Recognition Letters*, 25 (13), 1491–1500.

Bruzzone, L. and Marconcini, M., 2009. Toward the automatic updating of land-cover maps by a domain-adaptation SVM classifier and a circular validation strategy. *IEEE Transactions on Geoscience and Remote Sensing*, 47 (4), 1108–1122.

Butler, D., 2008. *Web data predict flu*. Nature Publishing Group.

Campbell, J.Y. and Shiller, R.J., 1988. Stock Prices, Earnings, and Expected Dividends. *The Journal of Finance*, 43 (3), 661–676.

Cao, Z. and Lin, C.-T., 2018. Inherent fuzzy entropy for the improvement of EEG complexity evaluation. *IEEE Transactions on Fuzzy Systems*, 26 (2), 1032–1035.

Casana, J., 2014. Regional-scale archaeological remote sensing in the age of big data: automated site discovery vs. brute force methods. *Advances in Archaeological Practice*, 2 (3), 222–233.

Chatfield, C., 2000. *Time-series forecasting*. CRC press.

Chen, H., Chiang, R.H., and Storey, V.C., 2012. Business intelligence and analytics: From big data to big impact. *MIS quarterly*, 36 (4).

Chen, J., Du, Y., Liu, L., Zhang, P., and Zhang, W., 2019. BBS Posts Time Series Analysis based on Sample Entropy and Deep Neural Networks. *Entropy*, 21 (1), 57.

Chen, J., Li, C., Li, Z., and Gold, C., 2001. A voronoi-based 9-intersection model for spatial relations. *International Journal of Geographical Information Science*, 15 (3), 201–220.

Chen, J., Liao, A., Cao, X., Chen, L., Chen, X., He, C., Han, G., Peng, S., Lu, M., Zhang, W., Tong, X., and Mills, J., 2015. Global land cover mapping at 30 m resolution: A POK-based operational approach. *ISPRS Journal of Photogrammetry and Remote Sensing*, 103, 7–27.

Chen, L., Tang, Y., Lv, M., and Chen, G., 2015. Partition-based range query for uncertain trajectories in road networks. *GeoInformatica*, 19 (1), 61–84.

Chen, Y., Jiang, H., Li, C., Jia, X., and Ghamisi, P., 2016. Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 54 (10), 6232–6251.

Chen, Y. and Sun, K., 2014. Information measurement of classification maps and scale effects. In *IEEE Conference Anthology* (pp. 1-5). IEEE.

Ciepluch, B., Jacob, R., Mooney, P., and Winstanley, A.C., 2010. Comparison of the accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps. *In*: *Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resuorces and Enviromental Sciences 20-23rd July 2010*. University of Leicester, 337.

Clements, M.P. and Hendry, D.F., 2000. *Forecasting non-stationary economic time series*. MIT Press.

Cleveland, W.S., 1979. Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, 74 (368), 829–836.

Colwell, R.K., 1974. Predictability, constancy, and contingency of periodic phenomena. *Ecology*, 55 (5), 1148–1153.

Congalton, R.G. and Green, K., 2008. *Assessing the accuracy of remotely sensed data: principles and practices*. CRC press.

Costa, M., Goldberger, A.L., and Peng, C.-K., 2002. Multiscale entropy analysis of complex physiologic time series. *Physical review letters*, 89 (6), 068102.

Dalton, C.M. and Thatcher, J., 2015. Inflated granularity: Spatial "Big Data" and geodemographics. *Big Data & Society*, 2 (2), 2053951715601144.

Davis, M.A. and Palumbo, M.G., 2001. *A primer on the economics and time series econometrics of wealth effects*. Divisions of Research & Statistics and Monetary Affairs, Federal Reserve.

Demšar, U., Harris, P., Brunsdon, C., Fotheringham, A.S., and McLoone, S., 2013. Principal Component Analysis on Spatial Data: An Overview. *Annals of the Association of American Geographers*, 103 (1), 106–128.

Desclée, B., Bogaert, P., and Defourny, P., 2006. Forest change detection by statistical object-based method. *Remote Sensing of Environment*, 102 (1–2), 1–11.

Di Gregorio, A., 2005. *Land cover classification system: classification concepts and user manual: LCCS*. Food & Agriculture Org.

Dias, D. and Costa, L.H.M.K., 2018. CRAWDAD dataset coppe-ufrj/RioBuses (v. 2018-03-19).

Donges, N., 2018. Recurrent Neural Networks and LSTM [online]. *Towards Data Science*. Available from: https://towardsdatascience.com/recurrent-neural-networks-and-lstm-4b601dd822a5 [Accessed 29 Mar 2019].

Drăguţ, L., Tiede, D., and Levick, S.R., 2010. ESP: A tool to estimate scale parameter for multiresolution image segmentation of remotely sensed data. *International Journal of Geographical Information Science*, 24 (6), 859–871.

Egenhofer, M.J. and Sharma, J., 1993a. Assessing the consistency of complete and incomplete topological information. *Geographical Systems*, 1 (1), 47–68.

Egenhofer, M.J. and Sharma, J., 1993b. Topological relations between regions in ρ2 and Z2. *Advances in Spatial Databases*, 316–336.

Epple, B., 2006. Using a GPS-aided inertial system for coarse-pointing of free-space optical communication terminals. *In*: *Free-Space Laser Communications VI*.

Presented at the Free-Space Laser Communications VI, International Society for Optics and Photonics, 630418.

European Environment Agency, 2007. *CLC2006 technical guidelines*. EEA Technical report.

Fan, W. and Bifet, A., 2013. Mining big data: current status, and forecast to the future. *ACM sIGKDD Explorations Newsletter*, 14 (2), 1–5.

Fisher, P.F., 1999. Models of uncertainty in spatial data. *Geographical Information Systems: Principles, Techniques, Management and Applications*, 191–205.

Foell, S., Phithakkitnukoon, S., Kortuem, G., Veloso, M., and Bento, C., 2015. Predictability of public transport usage: A study of bus rides in Lisbon, Portugal. *IEEE Transactions on Intelligent Transportation Systems*, 16 (5), 2955–2960.

Fonte, C.C., Antoniou, V., Bastin, L., Estima, J., Arsanjani, J.J., Bayas, J.-C.L., See, L., and Vatseva, R., 2017. Assessing VGI data quality. *Mapping and the citizen sensor*, 137–163.

Foody, G.M., 2010. Assessing the accuracy of land cover change with imperfect ground reference data. *Remote Sensing of Environment*, 114 (10), 2271–2285.

Foody, G.M., 2012. Latent class modeling for site- and non-site-specific classification accuracy assessment without ground data. *IEEE Transactions on Geoscience and Remote Sensing*, 50 (7), 2827–2838.

Frank, A.U., 1992. Qualitative Spatial Reasoning about Distance and Directions in Geographic Space. *Journal of Visual Languages and Computing*, 3, 343–373.

Furtado, A.S., Alvares, L.O.C., Pelekis, N., Theodoridis, Y., and Bogorny, V., 2018. Unveiling movement uncertainty for robust trajectory similarity analysis. *International Journal of Geographical Information Science*, 32 (1), 140–168.

Gao, P., Zhang, H., and Li, Z., 2017. A hierarchy-based solution to calculate the configurational entropy of landscape gradients. *Landscape Ecology*, 32 (6), 1133–1146.

Garland, J., James, R., and Bradley, E., 2014. Model-free quantification of time-series predictability. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 90 (5), 1–15.

Gerhardt, B., Griffin, K., and Klemann, R., 2012. Unlocking value in the fragmented world of big data analytics. *Cisco Internet Business Solutions Group*.

Gobble, M.M., 2013. Big Data: The Next Big Thing in Innovation. *Research-Technology Management*, 56 (1), 64–67.

Gödel, K., 1931. Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatshefte für mathematik und physik*, 38 (1), 173–198.

Goodchild, M.F., 2018. Reimagining the history of GIS. *Annals of GIS*, 24 (1), 1–8.

Goodchild, M.F. and Gopal, S., 1989. *The accuracy of spatial databases*. CRC Press.

Goodchild, M.F. and Li, L., 2012. Assuring the quality of volunteered geographic information. *Spatial statistics*, 1, 110–120.

Goodchild, M.F., Shi, W., and Fisher, P., 2003. Visualisation of Uncertainty in Geographical data. *In*: *Spatial Data Quality*. CRC Press, 165–184.

Guptill, S.C., 2017. Uncertainty. *International Encyclopedia of Geography: People, the Earth, Environment and Technology*, 1–8.

Gutiérrez, J., Cardozo, O.D., and García-Palomares, J.C., 2011. Transit ridership forecasting at station level: an approach based on distance-decay weighted regression. *Journal of Transport Geography*, 19 (6), 1081–1092.

Haklay, M. and Weber, P., 2008. Openstreetmap: User-generated street maps. *IEEE Pervasive Computing*, 7 (4), 12–18.

Hamilton, J.D., 1994. *Time series analysis*. Princeton university press Princeton, NJ.

Hand, D.J., 2006. Data Mining. *Encyclopedia of Environmetrics*, 2.

Hashemi, P. and Abbaspour, R.A., 2015. Assessment of logical consistency in OpenStreetMap based on the spatial similarity concept. *In*: *Openstreetmap in giscience*. Springer, 19–36.

He, F., Gu, L., Wang, T., and Zhang, Z., 2017. The synthetic geo-ecological environmental evaluation of a coastal coal-mining city using spatiotemporal big data: a case study in Longkou, China. *Journal of Cleaner Production*, 142, 854–866.

Heisenberg, W., 1925. Quantum-theoretical re-interpretation of kinematic and mechanical relations. *Z. Phys*, 33, 879–893.

Herold, M., Stehman, S.V., Wulder, M., Olofsson, P., Foody, G.M., Herold, M., Stehman, S.V., Woodcock, C.E., and Wulder, M.A., 2014. Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of Environment*, 148, 42–57.

Hochreiter, S. and Schmidhuber, J., 1997. Long Short-Term Memory. *Neural Computation*, 9 (8), 1735–1780.

Hornsby, K. and Egenhofer, M.J., 2002. Modeling moving objects over multiple granularities. *Annals of Mathematics and Artificial Intelligence*, 36 (1–2), 177–194.

Hoteit, S., Secci, S., Sobolevsky, S., Ratti, C., and Pujolle, G., 2014. Estimating human trajectories and hotspots through mobile phone data. *Computer Networks*, 64, 296–307.

Huang, Q. and Wong, D.W.S., 2015. Modeling and Visualizing Regular Human Mobility Patterns with Uncertainty: An Example Using Twitter Data. *Annals of the Association of American Geographers*, 105 (6), 1179–1197.

Hyndman, R.J. and Athanasopoulos, G., 2018. *Forecasting: principles and practice*. OTexts.

Ihler, A., Hutchins, J., and Smyth, P., 2007. Learning to Detect Events with Markov-modulated Poisson Processes. *ACM Trans. Knowl. Discov. Data*, 1 (3).

IBM. Big Data & Analytics Hub Extracting Business Value from the 4 V's of Big Data [online]. Available from: http://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data.

ISO, 2013. ISO 19157-2013: Geographic information -- Data quality. *International Standard*.

Jansen, L.J.M. and Gregorio, A.D., 2000. Land Cover Classification System (LCCS): Classification Concepts and User Manual. *FAO*.

Jeung, H., Lu, H., Sathe, S., and Yiu, M.L., 2014. Managing Evolving Uncertainty in Trajectory Databases. *IEEE Transactions on Knowledge and Data Engineering*, 26 (7), 1692–1705.

Jiang, D., Huang, Y., Zhuang, D., Zhu, Y., Xu, X., and Ren, H., 2012. A Simple Semi-Automatic Approach for Land Cover Classification from Multispectral Remote Sensing Imagery. *PLoS ONE*, 7 (9).

Jones, C.B., 2014. *Geographical information systems and computer cartography*. Routledge.

Jose, V.D., 2014. Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance & Society*, 12 (2), 197–208.

Jost, L., 2006. Entropy and diversity. *Oikos*, 113 (2), 363–375.

Kamilaris, A., Kartakoullis, A., and Prenafeta-Boldú, F.X., 2017. A review on the practice of big data analysis in agriculture. *Computers and Electronics in Agriculture*, 143, 23–37.

Kitchin, R., 2014. Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, 1 (1), 205395171452848.

Klemettinen, M., Mannila, H., and Toivonen, H., 1997. A data mining methodology and its application to semi-automatic knowledge acquisition. *In*: *Database and Expert Systems Applications. 8th International Conference, DEXA'97. Proceedings*. IEEE, 670–677.

Kuijpers, B. and Othman, W., 2006. Trajectory databases: Data models, uncertainty and complete query languages. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 4353 LNCS, 224–238.

Kuijpers, B. and Othman, W., 2009. Modeling uncertainty of moving objects on road networks via space–time prisms. *International Journal of Geographical Information Science*, 23 (9), 1095–1117.

Kwan, M.-P., 2016. Algorithmic geographies: Big data, algorithmic uncertainty, and the production of geographic knowledge. *Annals of the American Association of Geographers*, 106 (2), 274–282.

Kwan, M.-P., Wang, J., Tyburski, M., Epstein, D.H., Kowalczyk, W.J., and Preston, K.L., 2019. Uncertainties in the geographic context of health behaviors: A study of substance users' exposure to psychosocial stress using GPS data.

International Journal of Geographical Information Science, 33 (6), 1176–1195.

Lee, J., Kao, H.-A., and Yang, S., 2014. Service innovation and smart analytics for industry 4.0 and big data environment. *Procedia Cirp*, 16, 3–8.

Lee, J.-G. and Kang, M., 2015. Geospatial Big Data: Challenges and Opportunities. *Big Data Research*, 2 (2), 74–81.

Leibovici, D.G., Claramunt, C., Le Guyader, D., and Brosset, D., 2014. Local and global spatio-temporal entropy indices based on distance ratios and co-occurences distributions. *International Journal of Geographical Information Science*, 28 (5), 29–41.

Leszczynski, A., 2015. Spatial big data and anxieties of control. *Environment and Planning D: Society and Space*, 33 (6), 965–984.

Li, D., 1986. Theory of Separability for two Different Model Errors and Its Applications in Photogrammetric Point Determinations. *In*: *Proc. Symp. of Comm. III, ISPRS*. Rovanieni,Finland.

Li, S., 2007. Combining topological and directional information for spatial reasoning. *IJCAI International Joint Conference on Artificial Intelligence*, 137, 435–440.

Li, S., Dragicevic, S., Castro, F.A., Sester, M., Winter, S., Coltekin, A., Pettit, C., Jiang, B., Haworth, J., and Stein, A., 2016. Geospatial big data handling theory and methods: A review and research challenges. *ISPRS journal of Photogrammetry and Remote Sensing*, 115, 119–133.

Li, Z. and Huang, P., 2002. Quantitative measures for spatial information of maps. *International Journal of Geographical Information Science*, 16 (7), 699–709.

Lillesand, T., Kiefer, R.W., and Chipman, J., 2014. *Remote sensing and image interpretation*. John Wiley & Sons.

Liu, H., 2013. Methods of Measuring the Spatial Information Content of a Map. *Acta Geodaetica et Cartographica Sinica*, 42 (4).

Liu, H., Deng, M., Fan, Z., and Xu, Z., 2013. An Approach to Measuring the Spatial Information Content of a Point-shaped Map. *Acta Geodaetica et Cartographica Sinica*, 42 (1).

Liu, H., Deng, M., He, Z., and Xu, Z., 2012. An Approach to Measuring the Spatial

Information Content of an Area Feature, 14 (6).

Liu, X., Biagioni, J., Eriksson, J., Wang, Y., Forman, G., and Zhu, Y., 2012. Mining large-scale, sparse GPS traces for map inference: comparison of approaches. *In*: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 669–677.

Liu, X., Huang, Q., and Gao, S., 2019. Exploring the uncertainty of activity zone detection using digital footprints with multi-scaled DBSCAN. *International Journal of Geographical Information Science*, 33 (6), 1196–1223.

Lohr, S., 2012. The age of big data. *New York Times*, 11 (2012).

Lopez, D., Gunasekaran, M., Murugan, B.S., Kaur, H., and Abbas, K.M., 2014. Spatial big data analytics of influenza epidemic in Vellore, India. *In*: *2014 IEEE international conference on big data (Big Data)*. IEEE, 19–24.

Lou, Y., Zhang, C., Zheng, Y., Xie, X., Wang, W., and Huang, Y., 2009. Map-matching for low-sampling-rate GPS trajectories. *In*: *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems*. ACM, 352–361.

Lu, B., Charlton, M., Brunsdon, C., and Harris, P., 2016. The Minkowski approach for choosing the distance metric in geographically weighted regression. *International Journal of Geographical Information Science*, 30 (2), 351–368.

Lu, X., Bengtsson, L., and Holme, P., 2012. Predictability of population displacement after the 2010 Haiti earthquake. *Proceedings of the National Academy of Sciences*, 109 (29), 11576–11581.

Ma, X., Zhang, J., Du, B., Ding, C., and Sun, L., 2019. Parallel Architecture of Convolutional Bi-Directional LSTM Neural Networks for Network-Wide Metro Ridership Prediction. *IEEE Transactions on Intelligent Transportation Systems*, 20 (6), 2278–2288.

Mandelbrot, B., 1967. How long is the coast of Britain? Statistical self-similarity and fractional dimension. *science*, 156 (3775), 636–638.

Marin, R.-C., Dobre, C., and Xhafa, F., 2014. A methodology for assessing the predictable behaviour of mobile users in wireless networks. *Concurrency and*

Computation: Practice and Experience, 26 (5), 1215–1230.

McAfee, A., Brynjolfsson, E., Davenport, T.H., Patil, D.J., and Barton, D., 2012. Big data: the management revolution. *Harvard business review*, 90 (10), 60–68.

McInnes, L. and Healy, J., 2017. Accelerated Hierarchical Density Clustering. *In*: *IEEE International Conference on Data Mining Workshops*. IEEE, 33–42.

De Milliano, S., 2017. GIS: Challenges and Trends in 2017. *GIS Professional*. 76, 10-13.

Mullen, W.F., Jackson, S.P., Croitoru, A., Crooks, A., Stefanidis, A., and Agouris, P., 2015. Assessing the impact of demographic characteristics on spatial error in volunteered geographic information features. *GeoJournal*, 80 (4), 587–605.

Musolesi, M. and Mascolo, C., 2006. Evaluating context information predictability for autonomic communication. *In*: *2006 International Symposium on a World of Wireless, Mobile and Multimedia Networks(WoWMoM'06)*. Presented at the 2006 International Symposium on a World of Wireless, Mobile and Multimedia Networks(WoWMoM'06), 5 pp. – 499.

Neis, P., Zielstra, D., and Zipf, A., 2013. Comparison of volunteered geographic information data contributions and community development for selected world regions. *Future Internet*, 5 (2), 282–300.

Neumann, J., 1994. The Topological Information Content of a Map An Attempt at a Rehabilitation of Information Theory in Cartography. *Cartographica*, 31 (1), 26–34.

Niedermayer, J., Züfle, A., Emrich, T., Renz, M., Mamoulis, N., Chen, L., and Kriegel, H.-P., 2013a. Probabilistic Nearest Neighbor Queries on Uncertain Moving Object Trajectories. *Pvldb*, 7 (3), 205–216.

Niedermayer, J., Züfle, A., Emrich, T., Renz, M., Mamoulis, N., Chen, L., and Kriegel, H.P., 2013b. Similarity search on uncertain spatio-temporal data. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8199 LNCS, 43–49.

Olofsson, P., Foody, G.M., Stehman, S. V., and Woodcock, C.E., 2013. Making better use of accuracy data in land change studies: Estimating accuracy and area and

quantifying uncertainty using stratified estimation. *Remote Sensing of Environment*, 129, 122–131.

Otto, S.A., 2019. How to normalize the RMSE [online]. *How to normalize the RMSE*. Available from: https://www.marinedatascience.co/blog/2019/01/07/normalizing-the-rmse/.

Oxera, 2013. What is the Economic Impact of Geo Services?, (January), 1–42.

Pfoser, D. and Jensen, C.S., 1999. Capturing the uncertainty of moving-object representations. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 1651, 111–131.

Pincus, S.M., 1991. Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences*, 88 (6), 2297–2301.

Planet Team, 2017. Planet Application Program Interface: In Space for Life on Earth. *San Francisco, CA*.

Prank, K., Nowlan, S.J., Harms, H.M., Kloppstech, M., Brabant, G., Hesch, R.-D., and Sejnowski, T.J., 1995. Time series prediction of plasma hormone concentration. Evidence for differences in predictability of parathyroid hormone secretion between osteoporotic patients and normal controls. *The Journal of clinical investigation*, 95 (6), 2910–2919.

Preis, T., Moat, H.S., and Stanley, H.E., 2013. Quantifying trading behavior in financial markets using Google Trends. *Scientific reports*, 3, 1684.

Prigogine, I., Prigogine, I., Physicien, C., Prigogine, I., Physicist, C., and Prigogine, I., 1961. *Introduction to thermodynamics of irreversible processes*. Interscience Publishers New York.

Qiu, X., Ren, Y., Suganthan, P.N., and Amaratunga, G.A., 2017. Empirical mode decomposition based ensemble deep learning for load demand time series forecasting. *Applied Soft Computing*, 54, 246–255.

Radoux, J. and Defourny, P., 2010. Automated Image-to-Map Discrepancy Detection using Iterative Trimming. *Photogrammetric Engineering and Remote Sensing*, 76 (2), 173–181.

Radoux, J., Lamarche, C., Van Bogaert, E., Bontemps, S., Brockmann, C., Defourny, P., Radoux, J., Lamarche, C., Van Bogaert, E., Bontemps, S., Brockmann, C., and Defourny, P., 2014. Automated Training Sample Extraction for Global Land Cover Mapping. *Remote Sensing*, 6 (5), 3965–3987.

Ranacher, P. and Rousell, A., 2013. *An adaptive sampling approach for trajectories based on the concept of error ellipses*. Verlag der Österreichischen Akademie der Wissenschaften.

Ranu, S., Deepak P, Telang, A.D., Deshpande, P., and Raghavan, S., 2015. Indexing and matching trajectories under inconsistent sampling rates. *In*: *2015 IEEE 31st International Conference on Data Engineering*. Presented at the 2015 IEEE 31st International Conference on Data Engineering, 999–1010.

Richman, J.S. and Moorman, J.R., 2000. Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology-Heart and Circulatory Physiology*, 278 (6), H2039–H2049.

Romero, A., Gatta, C., and Camps-Valls, G., 2015. Unsupervised deep feature extraction for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 54 (3), 1349–1362.

Scarpino, S.V. and Petri, G., 2019. On the predictability of infectious disease outbreaks. *Nature communications*, 10, 898.

Schwanen, T., 2018. Uncertainty in Contextual Effects on Mobility: An Exploration of Causality. *Annals of the American Association of Geographers*, 0 (0), 1–7.

Senaratne, H., Mobasheri, A., Ali, A.L., Capineri, C., and Haklay, M., 2017. A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science*, 31 (1), 139–167.

Shahid, R., Bertazzon, S., Knudtson, M.L., and Ghali, W.A., 2009. Comparison of distance measures in spatial analytical modeling for health service planning. *BMC health services research*, 9 (1), 200.

Shannon, C.E., 1948. A Mathematical Theory of Communication. *Bell Labs Technical Journal*, 27 (3), 379–423.

Shannon, C.E. and Weaver, W., 1949. *The Mathematical Theory of Communication*.

University of Illinois Press.

Shaoyi, W., Zhao, W., and Qingyun, D., 2007. A measurement method of geometrical information considering multi-level map feature. *SCIENCE OF SURVEYING AND MAPPING*, 32 (4), 60–62.

Shi, W., 1998. A generic statistical approach for modelling error of geometric features in gis. *International Journal of Geographical Information Science*, 12 (2), 131–143.

Shi, W., 2009. *Principles of modeling uncertainties in spatial data and spatial analyses*. CRC press.

Shi, W., Chen, J., Zhan, Q., and Shu, H., 2012. Reliable Spatial Analysis, *Geomatics and Information Science of Wuhan University*, 37 (8), 883-887, 991.

Shi, W., Chen, P., and Zhang, X., 2017. Reliability Analysis in Geographical Conditions Monitoring. *Acta Geodaetica et Cartographica Sinica*, 10, 1620–1626.

Shi, W. and Liu, W., 2000. A stochastic process-based model for the positional error of line segments in GIS. *International Journal of Geographical Information Science*, 14 (1), 51–66.

Shi, W., Wang, S., Li, D., and Wang, X., 2003. Uncertainty-based spatial data mining. *Proc. Asia GIS Assoc. Wuhan, China*, 124–135.

Shi, W., Zhang, A., and Webb, G.I., 2018. Mining significant crisp-fuzzy spatial association rules. *International Journal of Geographical Information Science*, 32 (6), 1247–1270.

Shi, W., Zhang, A., Zhou, X., and Zhang, M., 2018. Challenges and Prospects of Uncertainties in Spatial Big Data Analytics. *Annals of the American Association of Geographers*, 4452, 1–8.

Šliupaitė, A., Navickas, Z., and Vainoras, A., 2015. Evaluation of complexity of ECG parameters using sample entropy and Hankel matrix. *Elektronika ir Elektrotechnika*, 92 (4), 107–110.

Smith, G., Wieser, R., Goulding, J., and Barrack, D., 2014. A refined limit on the predictability of human mobility. *In*: *2014 IEEE International Conference on*

Pervasive Computing and Communications (PerCom). IEEE, 88–94.

Song, C., Qu, Z., Blumm, N., and Barabási, A.-L., 2010. Limits of predictability in human mobility. *Science*, 327 (5968), 1018–1021.

Stephens, D.W., 1993. Learning and behavioral ecology: incomplete information and environmental predictability. *In*: *Insect learning*. Springer, 195–218.

Strahler, A.H., Boschetti, L., Foody, G.M., Friedl, M. a., Hansen, M.C., Herold, M., Mayaux, P., Morisette, J.T., Stehman, S.V., and Woodcock, C.E., 2006. Global Land Cover Validation: Recommendations for Evaluation and Accuracy Assessment of Global Land Cover Maps. *European Communities, Luxembourg, 51(4).*

Sui, D., Elwood, S., and Goodchild, M., 2012. Crowdsourcing geographic knowledge: volunteered geographic information (VGI) in theory and practice. *Springer Science & Business Media.*

Sukhov, V., 1967. Information capacity of a map entropy. *Geodesy and Aerophotography*, 10, 212–215.

Székely, G.J. and Rizzo, M.L., 2014. Partial distance correlation with methods for dissimilarities. *The Annals of Statistics*, 42 (6), 2382–2412.

Taylor, S.J. and Letham, B., 2018. Forecasting at Scale. *The American Statistician*, 72 (1), 37–45.

Tobler, W.R., 1970. A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, 46, 234.

Tong, X. and Shi, W., 2010. Measuring positional error of circular curve features in Geographic information systems (GIS). *Computers and Geosciences*, 36 (7), 861–870.

Tong, X., Wang, Z., Xie, H., Liang, D., Jiang, Z., Li, J., and Li, J., 2011. Designing a two-rank acceptance sampling plan for quality inspection of geospatial data products. *Computers and Geosciences*, 37 (10), 1570–1583.

Trajcevski, G., Choudhary, A., Wolfson, O., Ye, L., and Li, G., 2010. Uncertain range queries for necklaces. *Proceedings - IEEE International Conference on Mobile Data Management*, 199–208.

Trajcevski, G., Wolfson, O., Hinrichs, K., and Chamberlain, S., 2004. Managing uncertainty in moving objects databases. *ACM Transactions on Database Systems (TODS)*, 29 (3), 463–507.

Vagropoulos, S.I., Chouliaras, G.I., Kardakos, E.G., Simoglou, C.K., and Bakirtzis, A.G., 2016. Comparison of SARIMAX, SARIMA, modified SARIMA and ANN-based models for short-term PV generation forecasting. *In*: *2016 IEEE International Energy Conference (ENERGYCON)*. IEEE, 1–6.

Verbesselt, J., Hyndman, R., Newnham, G., and Culvenor, D., 2010. Detecting trend and seasonal changes in satellite image time series. *Remote sensing of Environment*, 114 (1), 106–115.

Walker, S.J., 2014. Big Data: A Revolution That Will Transform How We Live, Work, and Think. *International Journal of Advertising*, 33 (1), 181–183.

Wan, N., Lin Kan, G., and Wilson, G., 2017. Addressing location uncertainties in GPS-based activity monitoring: A methodological framework. *Transactions in GIS*, 21 (4), 764–781.

Wang, J., 2011. *Advances in Cartography and Geographic Information Engineering*. Surveying And Mapping Press.

Wang, L. and Zhang, S., 2014. Incorporation of texture information in a SVM method for classifying salt cedar in western China. *Remote Sensing Letters*, 5 (6), 501–510.

Wang, S., Shi, W., Yuan, H., and Chen, G., 2005. Attribute Uncertainty in GIS Data. *Fuzzy Systems and Knowledge Discovery*, 614–623.

Wang, W., Chau, K., Xu, D., and Chen, X.-Y., 2015. Improving forecasting accuracy of annual runoff time series using ARIMA based on EEMD decomposition. *Water Resources Management*, 29 (8), 2655–2675.

Wang, Z., 1990. *Principles of photogrammetry:(with remote sensing)*. Press of Wuhan Technical University of Surveying and Mapping.

Winters, P.R., 1960. Forecasting sales by exponentially weighted moving averages. *Management science*, 6 (3), 324–342.

Wolfert, S., Ge, L., Verdouw, C., and Bogaardt, M.-J., 2017. Big data in smart farming–

a review. *Agricultural Systems*, 153, 69–80.

Wu, X., Zhu, X., Wu, G.-Q., and Ding, W., 2013. Data mining with big data. *IEEE transactions on knowledge and data engineering*, 26 (1), 97–107.

Xu, T., Xu, X., Hu, Y., and Li, X., 2017. An entropy-based approach for evaluating travel time predictability based on vehicle trajectory data. *Entropy*, 19 (4), 165.

Yan, X.-Y., Zhao, C., Fan, Y., Di, Z., and Wang, W.-X., 2014. Universal predictability of mobility patterns in cities. *Journal of The Royal Society Interface*, 11 (100), 20140834.

Yao, H., Wu, F., Ke, J., Tang, X., Jia, Y., Lu, S., Gong, P., Ye, J., and Li, Z., 2018. Deep multi-view spatial-temporal network for taxi demand prediction. *In*: *Thirty-Second AAAI Conference on Artificial Intelligence*.

Zhang, A., Shi, W., and Webb, G.I., 2016. Mining significant association rules from uncertain data. *Data Mining and Knowledge Discovery*, 30 (4), 928–963.

Zhang, G. and Hsu, L.-T., 2019. A New Path Planning Algorithm Using a GNSS Localization Error Map for UAVs in an Urban Area. *Journal of Intelligent & Robotic Systems*, 94 (1), 219–235.

Zhang, J. and Goodchild, M.F., 2002. *Uncertainty in geographical information*. CRC press.

Zhao, Z., Shaw, S.-L., Xu, Y., Lu, F., Chen, J., and Yin, L., 2016. Understanding the bias of call detail records in human mobility research. *International Journal of Geographical Information Science*, 30 (9), 1738–1762.

Zhao, Z., Shaw, S.-L., Yin, L., Fang, Z., Yang, X., Zhang, F., and Wu, S., 2019. The effect of temporal sampling intervals on typical human mobility indicators obtained from mobile phone location data. *International Journal of Geographical Information Science*, 33 (7), 1471–1495.

Zheng, K., Trajcevski, G., Zhou, X., and Scheuermann, P., 2011. Probabilistic range queries for uncertain trajectories on road networks. *In*: *Proceedings of the 14th International Conference on Extending Database Technology - EDBT/ICDT '11*. Presented at the the 14th International Conference, Uppsala, Sweden: ACM Press, 283.

Zheng, Y., 2015. Trajectory data mining: an overview. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 6 (3), 29.

Zheng, Y., Zhang, L., Xie, X., and Ma, W.-Y., 2009. Mining interesting locations and travel sequences from GPS trajectories. *In*: *Proceedings of the 18th international conference on World wide web*. ACM, 791–800.

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

Appendix A Derivation process of the approximated ellipse for BAEE model

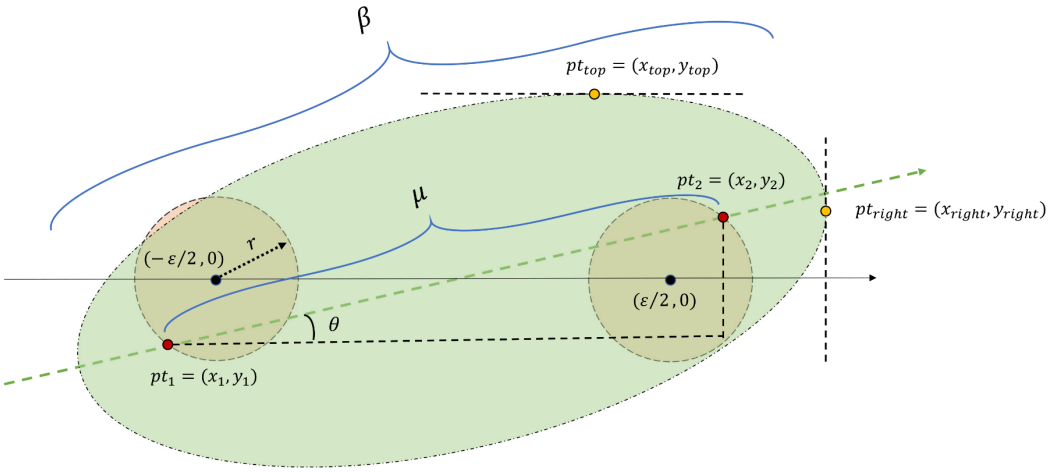


Figure A.1 Diagram for finding the maximum y or x of BAEE model.

Figure A.1 shows a general example that illustrates the calculation of extreme values of the BAEE model. Given two sampled points $(-\varepsilon/2, 0)$ and $(\varepsilon/2, 0)$, inferring that the theoretical BAEE model is an axisymmetric shape is easy. Therefore, determining the range for the arbitrary point (x, y) on BAEE boundary only requires the calculation of $\max(x)$ and $\max(y)$, whereas $\min(x) = -\max(x)$ and $\min(y) = -\max(y)$.

1) Limits of a rotated ellipse:

An ellipse centring at the origin with a rotation of θ can be expressed as:

$$\frac{(x\cos\theta - y\sin\theta)^2}{a^2} + \frac{(x\sin\theta + y\cos\theta)^2}{b^2} = 1, \tag{A-1}$$

where a and b denote the length of its semi-major and semi-minor axes, respectively. Through considerable algebra and tears, the limits of this ellipse can be easily obtained as follows:

$$\max(x) = \sqrt{a^2\cos^2\theta + b^2\sin^2\theta}, \max(y) = \sqrt{a^2\sin^2\theta + b^2\cos^2\theta}. \tag{A-2}$$

Therefore, the limits of an arbitrary ellipse with a rotation of θ and $pt_1 = (x_1, y_1)$ and $pt_2 = (x_2, y_2)$ as foci, such as the ellipse in Figure A.1, can be expressed as

$$\begin{aligned} \max(x) = x_{right} &= \sqrt{a^2 \cos^2\theta + b^2 \sin^2\theta} + \frac{x_1 + x_2}{2}, \\ \max(y) = y_{top} &= \sqrt{a^2 \sin^2\theta + b^2 \cos^2\theta} + \frac{y_1 + y_2}{2} \end{aligned} . \quad \text{(A-3)}$$

2) Possible locations of pt_1 and pt_2:

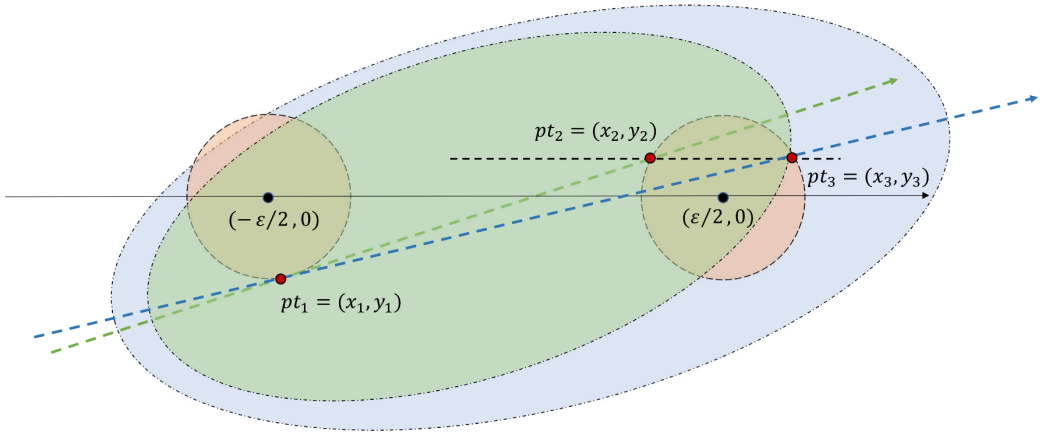


Figure A.2 Diagram for exploring the possible locations of point pair.

Given a certain $p_1 = (x_1, y_1)$, for any $p_2 = (x_2, y_2)$ on the left semi-circle as shown in Figure A.2, we can find a point $p_3 = (x_3, y_3)$ on the right semi-circle, where $y_3 = y_2$.

Note that the rotation θ can be represented by pt_1, pt_2; that is, $\cos\theta = \frac{x_2 - x_1}{u}, \sin\theta = \frac{y_2 - y_1}{u}$, and $a = \beta = c\mu/2$, $b = \sqrt{c^2 - 1} * u/2$. Therefore, Equation (A-2) can be simplified as

$$\begin{aligned} \max(x) = x_{right} &= \frac{1}{2}\sqrt{c^2 (x_2 - x_1)^2 + (c^2 - 1) * (y_2 - y_1)^2} + \frac{x_1 + x_2}{2}, \\ \max(y) = y_{top} &= \frac{1}{2}\sqrt{c^2 (y_2 - y_1)^2 + (c^2 - 1) * (x_2 - x_1)^2} + \frac{y_1 + y_2}{2} \end{aligned} . \quad \text{(A-4)}$$

Given $x_3 \geq x_2$, on the basis of Equation (A-4), we have $x_{right}(pt_1, pt_2) \leq x_{right}(pt_1, pt_3)$. For this reason, extrapolating the possible domain for pt_1, pt_2 corresponding to the maximum x value of the BAEE model is easy. The domain is shown in Figure A.3(a). Similarly, the domain for pt_1, pt_2 corresponding to the maximum y value of the BAEE model is shown in Figure A.3(b).

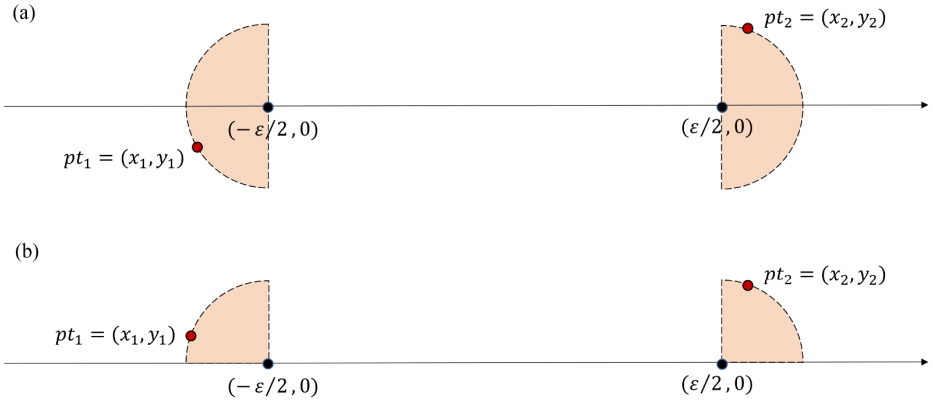


Figure A.3 Possible domain for pt_1, pt_2 corresponding to the limits of BAEE model.

3) Limits of BAEE (y-axis):

Let $E = \{e_0, e_1, e_2, \ldots\}$ denote an infinite set composed of the ellipses between all possible point pairs, where the maximum coordinate for $e \in E$ is expressed as x_e and y_e.

$$\forall k \in N, \exists e^y \in E \quad s.t. \quad y_{e^y} \geq y_{e_k} \tag{A-5}$$

Therefore, the computation of maximum y for theoretical BAEE model can be transformed into an objective function, which can be expressed as follows:

$$\max Z = \frac{1}{2}\sqrt{c^2(y_2 - y_1)^2 + (c^2 - 1) * (x_2 - x_1)^2} + \frac{y_1 + y_2}{2}$$
$$s.t. \begin{cases} (x_1 + \frac{\varepsilon}{2})^2 + y_1^2 = r^2 \\ (x_2 - \frac{\varepsilon}{2})^2 + y_2^2 = r^2 \end{cases} . \tag{A-6}$$

For the example in Figure A.1, the objective function Z can be rewritten as

$$Z=\frac{y_1+y_2}{2}+\frac{1}{2}\sqrt{c^2(y_1-y_2)^2+(c^2-1)(\sqrt{r^2-y_1^2}+\varepsilon+\sqrt{r^2-y_2^2})^2}\,. \quad \text{(A-7)}$$

The limits of Z appear at the stationary point (y_1^0, y_2^0), which should satisfy

$$Z_{y_1}(y_1^0,y_2^0)=\left.\frac{\partial Z}{\partial y_1}\right|_{\substack{y_1=y_1^0\\y_2=y_2^0}}=0 \quad \text{and} \quad Z_{y_2}(y_1^0,y_2^0)=\left.\frac{\partial Z}{\partial y_2}\right|_{\substack{y_1=y_1^0\\y_2=y_2^0}}=0\,. \quad \text{(A-8)}$$

After the differential, the partial derivative $Z_{y_1}(y_1^0, y_2^0)$, for instance, can be expressed as

$$Z_{y_1}(y_1^0,y_2^0)=\frac{c^2(y_1^0-y_2^0)-\dfrac{(c^2-1)*y_1^0*(\sqrt{r^2-(y_1^0)^2}+\varepsilon+\sqrt{r^2-(y_2^0)^2})}{\sqrt{r^2-(y_1^0)^2}}}{2\sqrt{(c^2-1)*(\sqrt{r^2-(y_1^0)^2}+\varepsilon+\sqrt{r^2-(y_2^0)^2})^2+c^2(y_1^0-y_2^0)^2}}+1\,. \quad \text{(A-9)}$$

Similarly, we can compute $Z_{y_2}(y_1^0, y_2^0)$ as follows:

$$Z_{y_2}(y_1^0,y_2^0)=\frac{c^2(y_2^0-y_1^0)-\dfrac{(c^2-1)*y_2^0*(\sqrt{r^2-(y_1^0)^2}+\varepsilon+\sqrt{r^2-(y_2^0)^2})}{\sqrt{r^2-(y_2^0)^2}}}{2\sqrt{(c^2-1)*(\sqrt{r^2-(y_1^0)^2}+\varepsilon+\sqrt{r^2-(y_2^0)^2})^2+c^2(y_1^0-y_2^0)^2}}+1\,. \quad \text{(A-10)}$$

Combining Equation (A-8)–(A-10) and after considerable algebra, we can obtain

$$y_1^0=y_2^0=\frac{r}{c} \quad \text{and} \quad Z_{\max}=r*c+\frac{\varepsilon}{2}\sqrt{c^2-1}\,. \quad \text{(A-11)}$$

4) Limits of BAEE (x-axis):

Similar to the derivation process in calculating the maximum y, the objective function Z can be written as:

$$Z=\frac{\sqrt{r^2-y_2^2}-\sqrt{r^2-y_1^2}}{2}+\frac{1}{2}\sqrt{(c^2-1)*(y_1-y_2)^2+c^2(\sqrt{r^2-y_1^2}+\varepsilon+\sqrt{r^2-y_2^2})^2}\,. \quad \text{(A-12)}$$

The partial derivative $Z_{y_1}(y_1^0, y_2^0)$ can be expressed as

$$Z_{y_1}(y_1^0,y_2^0)=\frac{(c^2-1)*(y_1^0-y_2^0)-\dfrac{c^2y_1^0*(\sqrt{r^2-(y_1^0)^2}+\varepsilon+\sqrt{r^2-(y_2^0)^2})}{\sqrt{r^2-(y_1^0)^2}}}{2\sqrt{c^2*(\sqrt{r^2-(y_1^0)^2}+\varepsilon+\sqrt{r^2-(y_2^0)^2})^2+(c^2-1)*(y_1^0-y_2^0)^2}}+\frac{y_1^0}{2\sqrt{r^2-(y_1^0)^2}}\,. \quad \text{(A-13)}$$

The partial derivative $Z_{y_2}(y_1^0, y_2^0)$ can be written as

$$Z_{y_2}(y_1^0, y_2^0) = \frac{(c^2\text{-}1)*(y_2^0\text{-}y_1^0) - \frac{c^2 y_2^0 * (\sqrt{r^2-(y_1^0)^2} + \varepsilon + \sqrt{r^2-(y_2^0)^2})}{\sqrt{r^2-(y_2^0)^2}}}{2\sqrt{c^2*(\sqrt{r^2-(y_1^0)^2} + \varepsilon + \sqrt{r^2-(y_2^0)^2})^2 + (c^2-1)*(y_1^0-y_2^0)^2}} - \frac{y_2^0}{2\sqrt{r^2-(y_2^0)^2}}. \quad \text{(A-14)}$$

Equating Equations (A-13) and (A-14) to zero yields $y_1^0 = y_2^0 = 0$. Therefore, the maximum x can be calculated as $\mathrm{Z}_{max} = c * (\frac{\varepsilon}{2} + r)$.