LOCATION-BASED SOCIAL MEDIA FOR ACTIVITY SPACE MODELING

by

Xujiao Wang, M.S., B.S.

A dissertation submitted to the Graduate Council of
Texas State University in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
with a Major in Geographic Information Science
August 2019

Committee Members:

Yihong Yuan, Chair

Byron Gao

Yongmei Lu

Alexander Savelyev

FAIR USE AND AUTHOR’S PERMISSION STATEMENT

Fair Use

Duplication Permission

ACKNOWLEDGEMENTS

First, I would specifically like to heartfeltly thank my advisor Dr. Yihong Yuan for her infinite patience, parental encouragement, and tremendous support during my 4-year doctoral study. When struggling with all kinds of difficulties, my perseverance would not be sustained without her encouragement. I am grateful for her dedication in teaching me how to conduct research and write scientific papers. This dissertation could not have been done without her expert guidance. Moreover, Dr. Yuan always inspired me by her professionalism and diligence. Her passionate attitude has motivated me to be a responsible scholar and human being. I greatly appreciate all her sincere advice on my studies and life. What I have learned from her will be a great treasure for my future career.

I would also like to sincerely thank all my committee members: Dr. Yongmei Lu, Dr. Byron Gao, and Dr. Alexander Savelyev. I am grateful for their insightful comments and dedicated time and efforts. Specifically, Dr. Lu provided me with valuable ideas on spatial-temporal study for my dissertation research. Special thanks go to Dr. Byron Gao for his efforts of commenting on the methodology of my proposal and dissertation. The research materials he provided have inspired me with great research thoughts. I sincerely thank Dr. Alexander Savelyev who has encouraged me to think critically in developing research questions. I am grateful to have had an opportunity to work with him as his research and teaching assistant. All of my committee members helped me tremendously in this dissertation.

I would like to give my gratitude to the professors in the department of geography. I thank Dr. F. Benjamin Zhan, Dr. Edwin Chow, Dr. Ron Hagelman, Dr. Richard Earl and Dr. Injeong Jo for their kind help and support. Many thanks to the staff in the department of geography. Allison Glass-Smith, Angelika Wahl, Pat Hell-Jones, Joyce Wilkerson, Charles Robinson, Stella LoPachin, and Dan Hemenway were greatly helpful throughout my doctoral studies.

I also thank all my dear friends herein: Jin He, Ping Boggess, Xiaowen Cui, Fuchuan Yang, Niaz Morshed, Jannatul Morshed, Ugo Frank Umeokafor, Yunuen Reygadas-Langarica, Francesco Zignol, Alisa Hartsell, Shadi Maleki, Milad Mohammadalizadehkorde, Jennifer Villa, Tasnuva Udita, Christina Lopez, David Mills, Molly Miranker, Joshua Hodge, Zahra Ghaffari, Mark Deka, David Szpakowski, and Brandy Clark for their warm encouragement and sincere help.

More importantly, I would like to thank my best friend and husband Guixing Wei for "pushing" me all the time. He is always my biggest motivation to move forward fearlessly. I thank all my family, Mom, Dad, my mother-in-law and father-in-law, and my 9-month-old daughter Mia; thank you all for being my biggest support!

Finally, I would like to thank the Texas State Geography department, the Graduate College, the University Student Government, and Dr. F. Benjamin Zhan for the financial support to my doctoral studies.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

ABSTRACT

Human activity research is rooted in the study of modeling the patterns of human activities in space and time. Previous studies have made prevalent progress in the theories, methods, and applications of human activity analysis. Among these studies, human activity space modeling has been a crucial topic in studying the spatial distribution of individual behaviors. Human activity space modeling aims to understand and solve various problems driven by human activities, such as urban expansion and traffic congestion in the process of urbanization. Many commonly used activity models in computational physics and computer science are constructed at an abstract and generic level. However, individual activities vary over space and time; it is therefore imperative to account for spatial-temporal dynamics and variations for activity space modeling at an individual-level.

Compared to traditional data sources that are costly and time-consuming to collect, the development of location-based social media (LBSM) has provided more flexibility for researchers regarding where, when, and how to collect information about human activity behaviors. Studies utilizing LBSM to analyze human activity patterns have grown rapidly. However, there is a lack of understanding about the morphology and the internal structure of activity space extracted from LBSM datasets. In addition, many studies lack effectiveness tests about how reliable LBSM data can be used to explain human activity space. To this end, this study explores the effectiveness of representing activity space from an individual perspective when using LBSM data from three Chinese cities (i.e., Beijing, Shanghai, and Guangzhou). The two objectives of this dissertation are summarized as follows:

First, due to the lack of effectiveness testing in deriving human movement from LBSM data, this study tests the effectiveness of intra-individual indicators in modeling activity spaces from LBSM data. We evaluate how data collection durations and the choice of indicators affect the reliability of intra-individual activity space modeling. We use a linear regression model with the logarithmic transformation to approximate how the magnitude of four external morphology features and three internal structure characteristics changes with different data collection durations – from 1 month to 12 months. The results demonstrate that as the data collection duration increases, the magnitude of all defined indicators approaches a steady point; however, there are also outlier users who exhibit distinct patterns. It provides a useful reference to explore the balance point between data effectiveness and appropriate sample size from the LBSM database on empirical analysis.

Second, little research was conducted to test the effectiveness of inter-individual models in comparing the internal structure of individual activity spaces based on unevenly distributed data. To fill this gap, this dissertation investigated how different models perform in identifying inter-individual similarities between LBSM users. We first clustered LBSM check-ins based on the density-based spatial clustering of applications with noise (DBSCAN). Appropriate clustering parameters are chosen with the help of the elbow method. We then import those clustered activities into a vector space model (VSM) and a spatial-temporal vector space model (ST-VSM). The former only considers the spatial locations of the check-ins, whereas the latter is also determined by the time period (i.e., morning, afternoon, and night) of the check-ins. We then measure LBSM user activity similarities by applying an extended cosine similarity analysis. The results successfully captured spatial-temporal activity similarities between LBSM users.

In conclusion, this study evaluated the effectiveness of LBSM for activity space modeling. Here we define “effectiveness” as the stability of activity space indicators with different amounts of data used. There are two contributions to the study of activity space modeling. On the one hand, this study explores the effectiveness of LBSM in modeling intra-individual activity space. The results of the effectiveness test demonstrated how data collection duration impacts the magnitude of different activity space indicators. As the data size increases, the magnitude of four external and three internal indicators all approach a steady point in three cities. It provided a useful reference to explore a balance point between effective indicators and the appropriate sample sizes from LBSM data. The indicators and methods used in this study can also be applied to other social media platforms to test their stability and extensibility. On the other hand, it provides a robust method to measure individuals’ spatial-temporal similarities based on LBSM data. We conducted an analysis to evaluate the effectiveness of different models in measuring the inter-individual similarity between LBSM users based on their unevenly distributed check-ins. The results indicated that the similarity measurement is effective in discovering the spatial-temporal similarity between LBSM users. This extended similarity measurement provided a more robust method to measure users’ activity similarities based on low-resolution LBSM data.

To sum up, this study generated valuable results in evaluating the effectiveness of LBSM for activity space modeling. The effectiveness tests on both intra-individual indicators and inter-individual similarity measures offer a new perspective on examining the performance of LBSM data in human activity space modeling. In addition, we also explored the activity patterns of the three largest cities in a rapidly developing country. The extracted activity patterns and outliers provided valuable input for urban planners and policy makers to understand the dynamics of urban residents in three densely populated Chinese cities. We foresee that this research will enhance the understanding of applying LBSM data to human activity studies and other widely applicable areas of geography, such as transportation, urban planning, and location-based services.

1. INTRODUCTION

1.1 Background

Jones, Koppelman and Orfueil (1990) defined activity analysis as a framework for analyzing daily or multi-day travel behaviors and deriving differences in lifestyles and activity between people. Among all the activity analysis research, activity space modeling has been a crucial topic in studying the spatial distribution of individuals' activity behaviors (Yuan, Raubal and Liu 2012). Previous studies have defined activity space as local areas within which people travel during their daily activities (Golledge and Stimson 1997). Activity spaces consist of locations and areas visited to satisfy the basic needs of people's everyday life. More specifically, individuals often visit a subset of a limited number of activity locations repeatedly (Schönfelder and Axhausen 2004, Gonzalez, Hidalgo and Barabasi 2008) - these locations as well as the travels between and around these locations form an individual activity space. Researchers have focused on both the morphology and the internal structure of human activity space. The former measures its basic characteristics (e.g., size, shape, etc.), and the latter emphasizes the reasons for which an activity space forms (e.g., regularly visited locations) (Golledge and Stimson 1997, Schönfelder and Axhausen 2002).

Traditional human activity space analysis often relies on travel surveys and questionnaires as data sources. However, collecting such data can be costly, time-consuming, and it is hard to acquire a sufficient sample set in a large spatial environment (Yuan and Wang 2018, Axhausen et al. 2002, Hasan, Zhan and Ukkusuri 2013). Meanwhile, the past few decades have witnessed the increasing availability of mobile devices with location sensors (e.g., the Global Positioning System (GPS)) and the booming of user-friendly client-side social networking applications (apps) (Hasan et al. 2013, Lane et al. 2010, Sakaki, Okazaki and Matsuo 2010, Stefanidis, Crooks and Radzikowski 2013). These new technologies have provided more flexibility regarding where, when, and how people connected to each other (Carrasco et al. 2008, Eagle, Pentland and Lazer 2009, Liben-Nowell et al. 2005). In the meantime, they also generate valuable datasets for researchers in the field of activity studies (Ahas et al. 2015, Doran et al. 2016, Lee et al. 2016, Resch 2013, Wu et al. 2014, Xu et al. 2015). Among these datasets, location-based social media (LBSM) is a popular and effective data source that attracts individuals to share their daily activities, whereabouts, and therefore provides abundant information about individuals' daily activities. The origins of the term "LBSM" is from the term "locative media", which uses geographical location through mobile devices in the social network (GSM Association 2003, Quercia et al. 2010, Steiniger et al. 2008, Wang, Min and Yi 2008). Compared to other types of data, LBSM not only provides non-spatial information, such as individuals' thoughts and emotions, but also generates geo-referenced data like users' locations, which can be related back to the points of interest (POIs). LBSM records social activities and interactions that happened in real locations (Varnelis and Friedberg 2008, Tuters and Varnelis 2006, Thielmann 2010, Sui and Goodchild 2011, Hemment 2006). It also strengthens the role of social media as a proxy for understanding human behaviors and complex social dynamics in geographic spaces (Cao et al. 2015). Hence, LBSM data offers various opportunities for researchers to explore and understand human activity patterns from both the urban and the individual perspectives (Liben-Nowell et al. 2005, Yuan and Medel 2016). From the urban perspective, researchers focused on how user activities exhibit universal properties and interact with urban structure and activities

(Bawa-Cavia 2011, Cranshaw et al. 2012, Mohammady and Culotta 2014, Phithakkitnukoon et al. 2010, Cho, Myers and Leskovec 2011, Hasan et al. 2013, Malleson and Birkin 2014, Wu et al. 2014). From the individual perspective, LBSM is particularly suitable for modeling individual-level patterns, such as activity scheduling, social network structure, and location prediction (Cho et al. 2011, Hasan et al. 2013, Malleson and Birkin 2014, Bawa-Cavia 2011, Calabrese et al. 2013, Gonzalez et al. 2008, Wu et al. 2014).

1.2 Problem Statement and Research Aims

Although human activity space modeling has been extensively studied, there are some limitations in the literature on analyzing human activity space from LBSM. In general, modeling human activity space based on LBSM data is helpful for understanding our socioeconomic environments (Chapin 1974, Liu et al. 2015, Aggarwal and Ryoo 2011). Although many studies have attempted to classify neighborhoods (Cranshaw et al. 2012) or extract activity anchor points (e.g., “home”, “work”) (Qu and Zhang 2013) from LBSM, there is a lack of understanding about the morphology and the internal structure of activity space from LBSM datasets (Malleson and Birkin 2014). In addition, many studies lack effectiveness tests to determine how reliable LBSM data can be utilized for presenting human activity space features (Brockmann, Hufnagel and Geisel 2006, Gonzalez et al. 2008). Here we use “effectiveness” as the stability of activity space indicators and measurements with different data input and algorithms in quantitative analysis. Effectiveness analysis is necessary for exploring human activity because it can test whether the modeling of activity space is robust with respect to various factors such as data quality, selected indicators, measurements, etc. Previously, there has not been sufficient research to evaluate the effectiveness of LBSM data sampling in deriving active space indicators or testing the effectiveness of identifying user similarities based on highly dispersed check-ins in LBSM-based studies. These two aspects are the main components of this dissertation.

Therefore, the effectiveness analysis in this dissertation can be conducted from the following two perspectives: intra-individual studies and inter-individual studies (Schönfelder and Axhausen 2003, Lee et al. 2016). The former evaluates the effectiveness of intra-individual activity space variability by using LBSM data. The latter focuses on addressing inter-individual activity space differences by comparing LBSM users' activity similarities. The specific research aims are listed as follows:

Research Aim 1: Test the effectiveness of intra-individual indicators in modeling activity space from LBSM data. This research aims to test the effectiveness of representing human activity space from both the indicator/measurement and the data input perspectives. This study evaluates how the choice of intra-individual indicators affects the effectiveness of modeling the external morphology and internal structures of individual activity spaces. We also test how LBSM data issues, such as low sampling resolution and uneven check-in frequency, may impact the effectiveness of different activity space indicators and the choice of data collection duration in the experiment design. This research task not only reviews the robustness of activity space indicators when applied to LBSM data, but also proposes a data processing strategy that can be extended to other datasets and activity space measures to help researchers optimize their research design.

Research Aim 2: Test the effectiveness of inter-individual models and measurements in comparing activity space patterns. In addition to intra-individual indicators, this research also evaluates the effectiveness of inter-individual models in presenting the similarities among activity spaces. First, we partition user locations into clusters according to their geographic coordinates and timestamps based on a density-based clustering method and vector-based method. Second, we measure the similarity of spatial activities between LBSM users based on the partitioned density surface. We then investigate their temporal activity patterns at different times of the day according to their check-in timestamps (i.e., morning, afternoon and night). This study explores the differences between similarity measures when applied to low resolution LBSM data.

1.3 Significance

The main contribution of this research is to test the effectiveness of LBSM data in modeling individual activity spaces. Activity space modeling is beneficial for discovering human activity characteristics.

First, this study measures the characteristics of each LBSM user's activity space derived from LBSM data. More specifically, it reveals the effectiveness of four external and three internal activity space indicators in measuring user activity spaces based on LBSM data. Moreover, this study tests how different amounts of check-in data affect the calculation of these activity space indicators in three Chinese cities. The results provide a useful reference for future experimental design in human activity modeling.

Second, this study detects inter-individual similarities based on low-resolution LBSM data. To understand the distinctions between individual activity spaces, it is important to explore the differences in the point patterns that form the activity spaces. In addition to location data, LBSM check-ins also provide the temporal signature of user activities. We then proposed similarity measurements for discovering similar users in social media datasets by taking into account both check-in locations and timestamps.

This study offers a data processing strategy to understand human activity spaces and patterns by analyzing their posts on LBSM. The findings will provide a better understanding of the effectiveness of using LBSM data for both intra-individuals' activity space analysis and inter-individuals' pattern detection. The proposed models have important implications for many activity space-related applications such as land-use planning, transportation design, and community detection, etc. In addition, discovering similar social media users is beneficial for many applications based on social media analytics, such as social community detection, friendship analysis, anomaly behavior detection and so on.

The remainder of this dissertation is organized as follows: Chapter 2 describes the data and study areas used in activity space modeling; Chapter 3 tests the effectiveness of activity space indicators and data collection durations for intra-individual activity space modeling; and Chapter 4 tests the effectiveness of models in measuring inter-individual activity space similarities. We conclude this dissertation and present directions for future work in Chapter 5.

2. DATA AND STUDY AREA

2.1 Data Collection

China provides an ideal data environment for studying an individual's activity space through social media data. The social media market has developed rapidly in China and attracts more and more users. Sina Weibo is a Chinese microblogging website launched by Sina Corporation on August 14th, 2009. Its monthly active users ("MAUs") reached 431 *million* in June 2018 according to its unaudited financial report from June 30, 2018 (CIW 2018). Such rapid expansion indicates that Weibo users can easily access the internet to post information about their activities.

For this study, we demonstrated the use of an inexpensive and easy-to-collect long-term Sina Weibo dataset to address the effectiveness in activity space modeling research. Our social media data have been obtained from April 2015 to March 2016 through the official Weibo application programming interface (API). We only utilized a few fields related to this study, such as the unique identifier (i.e., user account ID), the coordinates of check-in locations, and the timestamp of check-ins as our data attributes among all the fields extracted from LBSM data. Table 2.1 lists the field example we use in further research. We implemented the research by using individuals' check-in data in Beijing, Shanghai, and Guangzhou.

Table 2.1 Sample Records of Sina Weibo Check-ins.

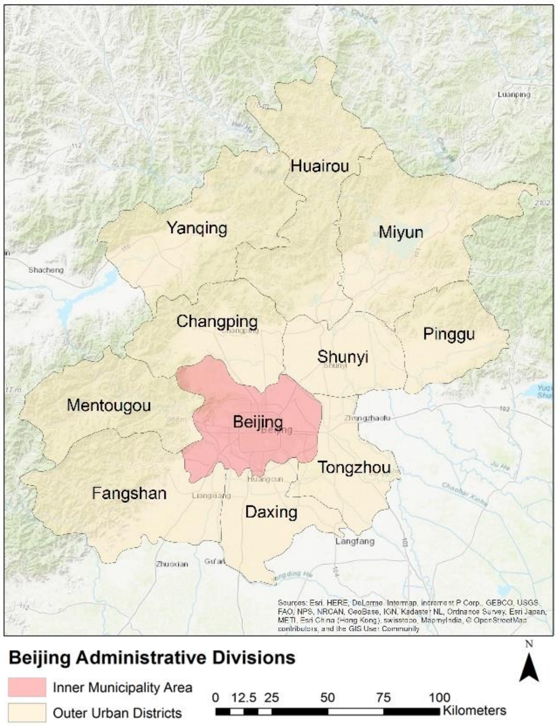
User ID	Timestamp	Longitude	Latitude
187811******	2015-06-25 05:51:53	116.599239	39.908899
520391******	2015-11-11 11:27:09	116.419662	40.090118

We grouped Weibo check-ins by counting the number of unique users and their total posting in monthly growth. It provides information about the activeness of users who uses the website on a monthly basis. Aside from a monthly basis, active users can be measured daily and weekly. However, daily or weekly frequency is not sufficient in quantity usage in this study due to the low-frequency check-ins for most social media users. The monthly period allows a longer and reasonable time collection period for most types of LBSM users.

2.2 Study Area

This dissertation uses the three most populated cities in China as the study areas. These three cities are Beijing, Shanghai, and Guangzhou. All of them are highly developed and populated cities in northern, eastern and southern metropolitan areas of China respectively.

Beijing, the capital of China, is located in the northern part of China. It is a densely-populated megacity with a total population of 21.73 million and an area of 16,810 km^2 in 2016 (National Bureau of Statistics of China 2017). Beijing currently comprises 16 administrative county-level subdivisions including one inner municipality area (i.e., the central area, defined as “Shixiaqu” in China) and ten outer urban districts (i.e., within the administrative boundary of the city but outside of the “Shixiaqu” area) (Figure 2.1). Hence, we separate the study areas into two parts for future analysis: the inner municipality area and the urban districts outside of the central municipality area but still within the administrative boundary (i.e., the outer urban districts) (Faber 2014).



Sources: Esri, HERE, DeLorme, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), swisstopo, MapmyIndia, © OpenStreetMap contributors, and the GIS User Community

Figure 2.1 Beijing Administrative Divisions.

Shanghai, the financial center of China, is located in the eastern part of China. Shanghai is one of the four municipalities directly under the control of the central government of China. It is one of the most populous cities in the world, with a population of more than 24.2 million within an area of 6,340 km^2 as of 2016 (National Bureau of Statistics of China 2017). It is also a transport hub, with the world's busiest container port. Here we focus on the Shanghai metropolitan area excluding the Chongming island, because the Chongming area is isolated from the rest of Shanghai and consists of three low-lying inhabited islands. Shanghai is divided into one inner municipality area and nine outer urban districts (Figure 2.2).

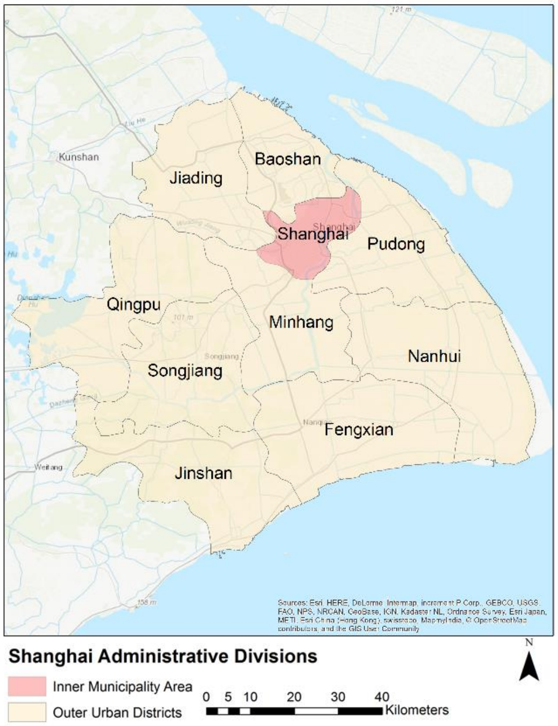

Baoshan
Jiading
Kunshan
Shanghai
Pudong
Qingpu
Minhang
Nanhui
Songjiang
Fengxian
Jinshan
Sources: Esri, HERE, DeLorme, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), swisstopo, MapmyIndia, © OpenStreetMap contributors, and the GIS User Community
N
Shanghai Administrative Divisions
Inner Municipality Area
Outer Urban Districts
0 5 10 20 30 40
Kilometers

Figure 2.2 Shanghai Administrative Divisions.

Guangzhou is the capital and most populous city of the Guangdong Province. It has a total population of 14.04 million within an area of 7,434 km^2 in 2016 (Guangzhou International 2016). It is located at the heart of the metropolitan area in southern China - an area that extends into the neighboring cities of Foshan, Dongguan, and Shenzhen, forming one of the largest urban agglomerations in the world. Guangzhou is a sub-provincial city. It has direct jurisdiction over one inner municipality area and five outer urban districts (Figure 2.3).

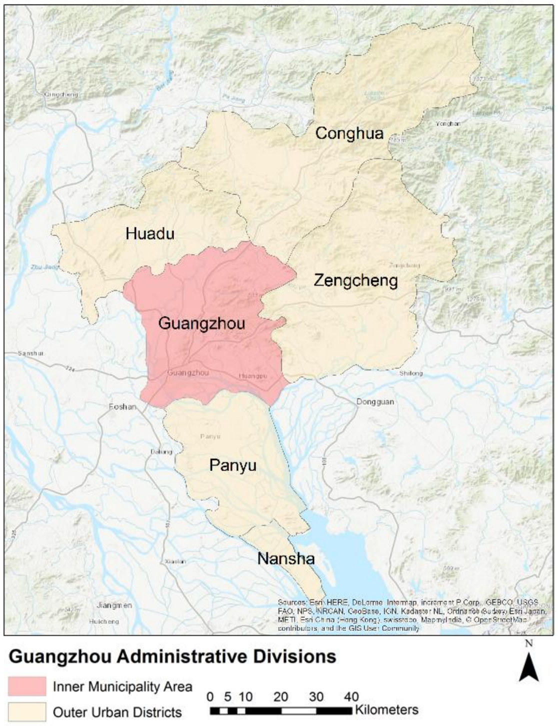

Conghua
Huadu
Zengcheng
Guangzhou
Panyu
Nansha
Foshan
Dongguan
Jiangmen
Sources: Esri, HERE, DeLorme, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), swisstopo, MapmyIndia, © OpenStreetMap contributors, and the GIS User Community
Guangzhou Administrative Divisions
Inner Municipality Area
Outer Urban Districts
0 5 10 20 30 40
Kilometers
N

Figure 2.3 Guangzhou Administrative Divisions.

We extracted Beijing, Shanghai, and Guangzhou administrative boundaries' data from the GADM database (GADM 2018) of version 2.8, November 2015. Note that we only included the population within the city limit instead of the entire metropolitan area.

2.3 Data Pre-processing

This study used the Sina Weibo check-ins to model individuals' activity spaces. Three datasets collected from Sina Weibo in Beijing, Shanghai, and Guangzhou are used to derive LBSM users' activity space indicators. Figure 2.4 illustrates the details of the data pre-processing procedure. We extracted 1.18 million geo-referenced Weibo posts (i.e., check-ins) for all three cities from April 2015 to March 2016 through the official Weibo API. Figure 2.5 shows Weibo check-in data's distribution, we can see the highest density of check-ins located in the center of the city in Beijing, Shanghai, and Guangzhou, China.

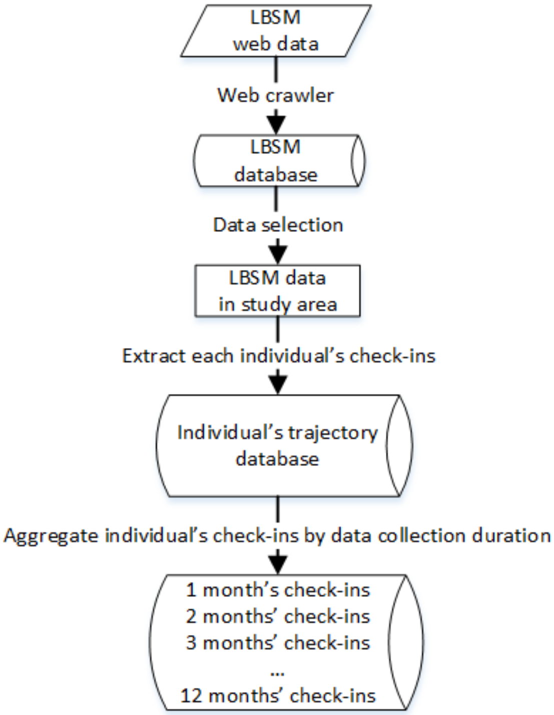
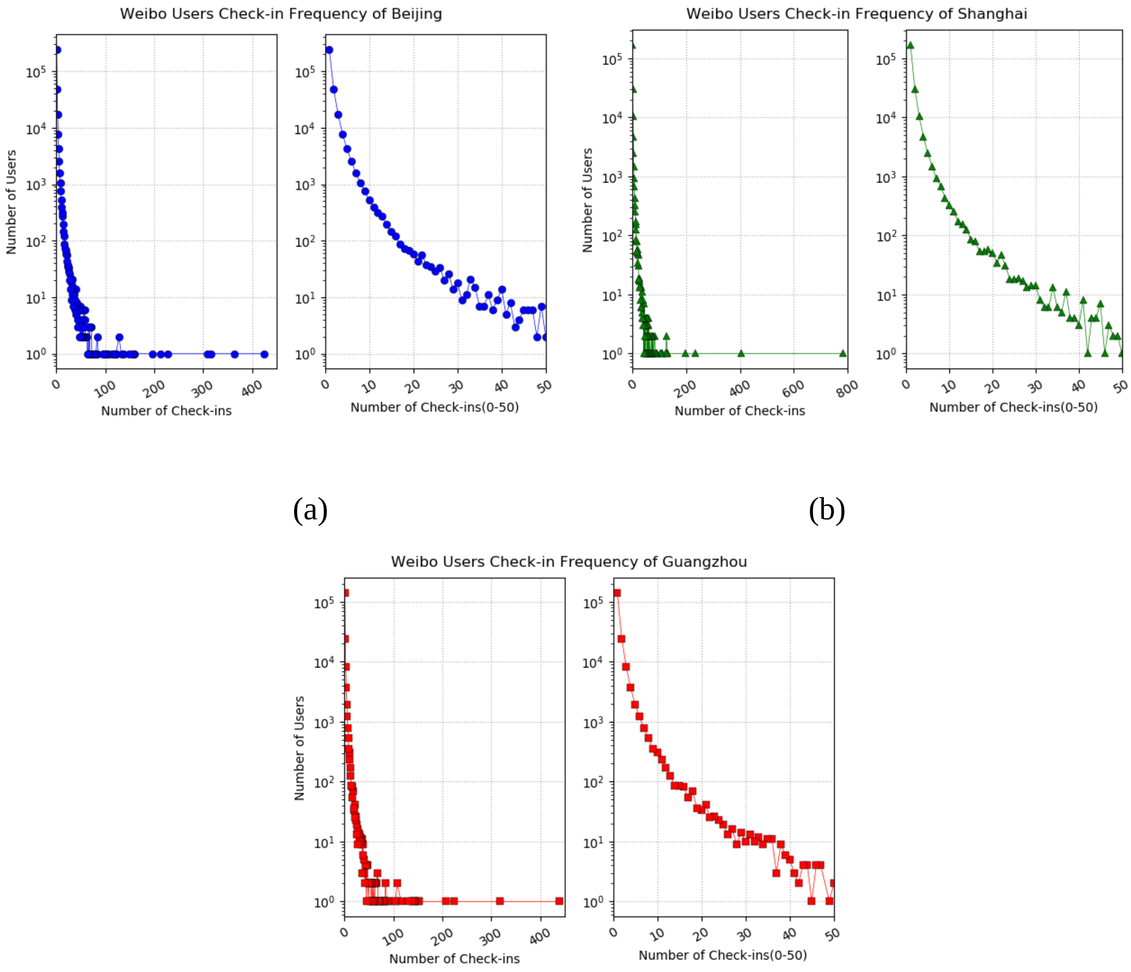


Figure 2 4 Workflow of Data Collection and Preparation.

Table 2.2 Data Summary before Filtering

Check-ins / City	Total check-ins	Total users	Average posting per user
Beijing	533,733	327,340	1.6
Shanghai	354,286	219,296	1.6
Guangzhou	293,737	181,207	1.6

We calculated the average posting per Weibo user in three study areas respectively. Based on the number of check-ins and users we collected from the Weibo streaming API, the result shows that the average posting per user is fewer than 2 check-ins, which is too sparse to show individual activity patterns (Table 2.2). In addition, in order to see the detailed patterns of check-in frequency of LBSM users, we summarized the frequency distribution of Weibo users in Figure 2.6. For each city, the left sub-figure shows the distribution of all the data, and the right sub-figure shows a "zoom-in" view of users with fewer than 50 check-ins from April 2015 to March 2016. The Y-axis counts the number of users who share the same check-in frequency, and the X-axis denotes the usage frequencies of LBSM.



(a) (b)

(c)

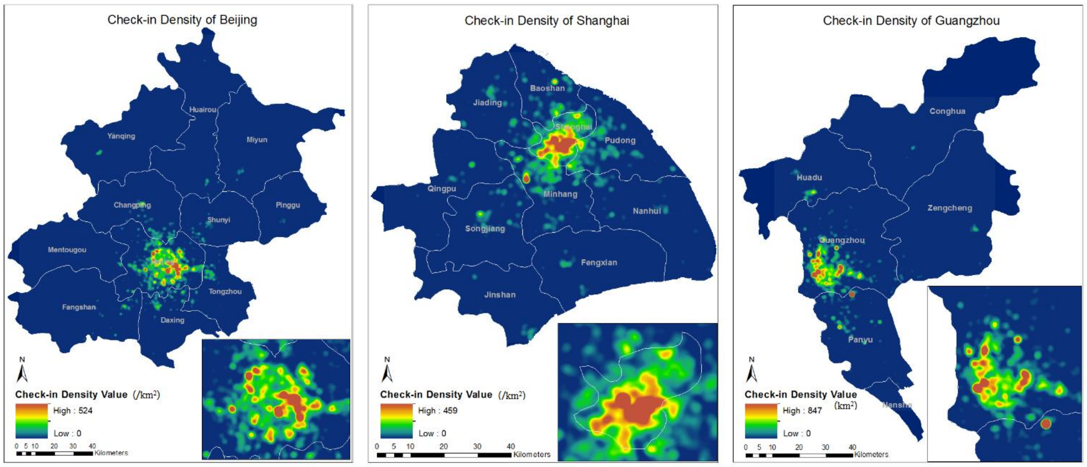
Figure 2.5 The Frequency Distribution of Weibo Check-in Data:

(a) Beijing; (b) Shanghai; and (c) Guangzhou.

Figure 2.5 shows a power-law distribution in which a few users post a lot and most posts are contributed by users who seldomly use Weibo. Due to the power law nature of social media usage, many of these posts are from inactive users. Considering the sparseness issue of LBSM data, this research focuses on the LBSM users who are active to ensure that we have reliable information to extract individual users' activity patterns. We eliminated users with fewer than 10 check-ins during the data collection period, as their activity spaces may be sensitive to outlier points. We choose 10 check-ins as a pre-defined threshold for this case study. This allows us to mitigate the data sparseness issue by removing potential short-term visitors and/or inactive LBSM users during the study period. The goal of this study is to demonstrate the feasibility of the methodology, and future research can explore how different thresholds may impact the results based on different LBSM sample data. Table 2.3 shows that the average posting per user after filtering reaches 18, which better describes individual user patterns.

Table 2.3 Data Summary after Filtering Process

Check-ins City	Check-ins after filtering	Users (whose check-ins>=10)	Average posting per user
Beijing	50,355	2,800	18
Shanghai	33,122	1,776	19
Guangzhou	30,608	1,639	19


Check-in Density of Beijing
Huairou
Yanqing
Miyun
Changping
Pinggu
Shunyi
Mentougou
Tongzhou
Fangshan
Daxing
N
Check-in Density Value (/km²)
High : 524
Low : 0
0 5 10 20 30 40 Kilometers
Check-in Density of Shanghai
Baoshan
Jiading
Shanghai
Pudong
Qingpu
Minhang
Nanhui
Songjiang
Fengxian
Jinshan
N
Check-in Density Value (/km²)
High : 459
Low : 0
0 5 10 20 30 40 Kilometers
Check-in Density of Guangzhou
Conghua
Huadu
Zengcheng
Guangzhou
Panyu
Nansha
N
Check-in Density Value (km²)
High : 847
Low : 0
0 5 10 20 30 40 Kilometers

(a) (b) (c)

Figure 2.6 Weibo Check-in Data Distribution and Density:

(a) Beijing; (b) Shanghai; and (c) Guangzhou.

Table 2.4 shows the percentage of check-ins in the inner municipality area and the outer urban districts. The check-ins within the municipality areas of Beijing and Guangzhou occupied more than 70% over the total check-ins in the areas. Meanwhile, in megacities like Beijing or Shanghai, many local residents live in outer urban districts and work in the municipality area, so their activity space goes beyond the inner municipality area (Na, Yanwei and Mei-Po 2015, Xu et al. 2015). This is particularly important for cities like Shanghai, where only 39% of check-ins are from the inner municipality area since Shanghai has a small urban center.

Table 2.4 Amount and Percentage of Check-ins in Inner and Outer Areas.

	Check-ins in inner area		Check-ins in outer areas	
	Amount	Percentage	Amount	Percentage
Beijing	36,014	71.52%	14,341	28.48%
Shanghai	12,927	39.03%	20,195	60.97%
Guangzhou	21,827	71.31%	8,781	28.69%

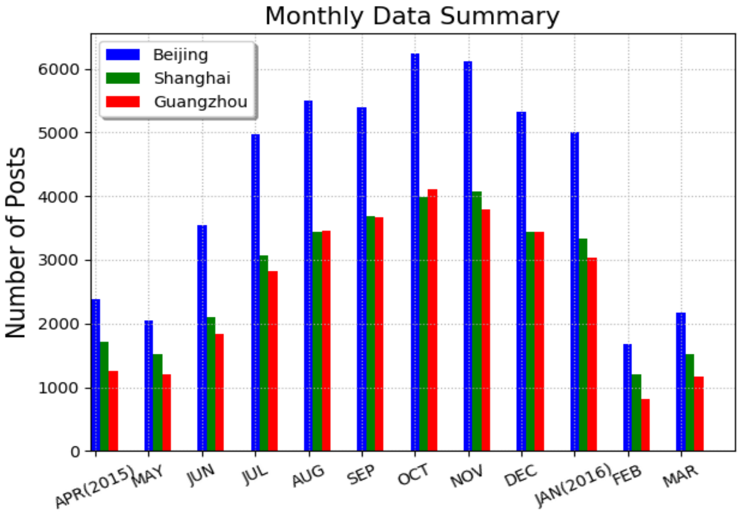


Figure 2.7 Monthly Check-in Data after Data Filtering.

Figure 2.7 shows the number of check-ins after data cleaning by removing inactive users who posts fewer than 10 times. From the monthly check-in summary, Weibo check-ins indicate a strong seasonal pattern in all three cities, where October and November are the most active months. This is potentially due to the 9-day national holiday (the National Day) in China towards the beginning of October, during which Chinese residents often spend leisure time with their family.

Table 2.5 Data Summary of Accumulated Check-ins.

Number of months (1-12)	Beijing	Shanghai	Guangzhou
1	2388	1709	1258
2	4430	3228	2462
3	7975	5335	4304
4	12942	8410	7136
5	18448	11858	10586
6	23839	15544	14250
7	30077	19533	18354
8	36189	23604	22152
9	41512	27048	25591
10	46512	30388	28625
11	48190	31593	29435
12	50355	33122	30608

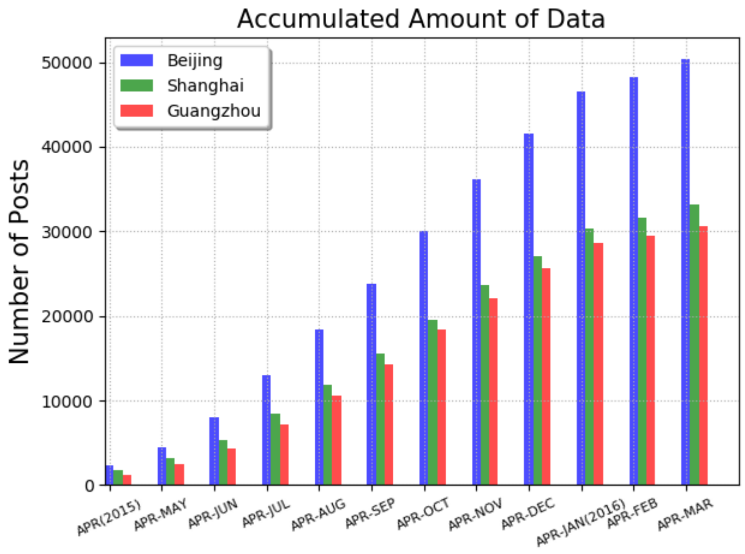


Figure 2.8 Accumulated Weibo Data in Study Areas.

Also, we display the magnitude of accumulated data collection (Table 2.5 and Figure 2.8). We aggregated data into 12 different sizes, from 1-month to 12-month data collection durations for each study area, to facilitate the effectiveness test on data collection durations (c.f., Section 3.3.2).

3. TEST THE EFFECTIVENESS OF INTRA-INDIVIDUAL ACTIVITY SPACE MODELING

3.1 Introduction

In the big data era, LBSM has been widely utilized as a supplement to traditional surveys in modeling human activity patterns (Sui and Goodchild 2011). However, there has not been sufficient studies to assess the reliability of these data in deriving human movement. Hence, determining the appropriate data size, duration, and sampling resolution are crucial for designing a statistically sound study. In fact, these factors are often determined arbitrarily when using LBSM to analyze activity patterns, and there has yet to be a systematic study of how users' activity spaces change with different sample sizes from LBSM. This chapter addresses evaluating appropriate activity space indicators and data sizes to achieve a balance between the details of information and computation efficiency/data collection costs.

This chapter evaluates how data collection duration and the choice of indicators affect the reliability of LBSM data in intra-individual activity space modeling. The effectiveness test is conducted based on the measurement of 7 activity space indicators and 12 LBSM data collection durations. Four of the activity space indicators are external morphology indicators which are minimum convex hull, alpha shape, standard deviational ellipse (SDE), and radius of gyration (ROG). The remaining three are internal structure indicators which are entropy, kernel density, and the minimum spanning trees (MST).

We use Weibo data as an example to illustrate how the magnitude of each activity space indicator changes with different LBSM data sizes. Besides, we estimate an optimal data size by applying a linear regression model with logarithmic transformation. It reflects the correlation between the change of indicator values and the amount of data used, and to approximate the limit values of indicators. This case study focuses on three Chinese cities (Beijing, Shanghai, and Guangzhou) and provides a useful reference to explore the balance point between data effectiveness and appropriate sample size from LBSM data.

3.2 Previous Work on Activity Space Modeling

Our understanding of human activity behaviors can be deeply enriched through observing and analyzing their activity spaces. Activity space is defined as the local areas that people travel within while performing their daily activities (Becker et al. 2013, Mazey 1981, Yuan and Raubal 2016). The components of an individual activity space include their travel trajectories, POIs, and the interactions with the environment (Golledge 1997, Lewin 1951).

As Golledge and Stimson (1997) pointed out, there are three determinants of activity space for a given individual: 1) the position of the individual home location; 2) regularly visited activity locations such as the work location, grocery stores, park, cinemas, etc.; and 3) travel between and around the pegs, such as the accessibility of public transport to regularly visited locations. Schönfelder and Axhausen (2016) extended this definition and identified six elements of activity space: home location, duration of residence, number of activity locations in the vicinity of home, trips within the neighborhood, mobility to and from frequently visited activity locations, and travels between the centers of daily life. Järv, Ahas and Witlox (2014) found modest monthly variation in the number of activity locations by observing individual activities for 12 consecutive months, whereas there were great variations in the sizes of individual activity spaces. Schönfelder and Axhausen (2003) discovered that the main factor deciding the scale of the activity spaces is the overall number of unique locations visited by individuals. Therefore, many previous studies concentrated on extracting activity anchor points and individual differences of visiting these points, as well as understanding the formation of activity spaces (Ahas et al. 2015, Long and Nelson 2013, Malleson and Birkin 2014, Phithakkitnukoon et al. 2010, Silm and Ahas 2014, Xu et al. 2015, Xu et al. 2016).

3.2.1 Usage of LBSM in Activity Space Studies

Traditional travel records are difficult to collect in the long term for a large group of participants (Dijst 1999, Fan and Khattak 2008, Kim and Ulfarsson 2015). Unlike traditional survey or individually-collected GPS data, LBSM datasets cover a large sample size and can easily be accessed by APIs, therefore they provide a rich resource for researchers to analyze human activity patterns. Nowadays, LBSM like Twitter and Weibo become particularly promising data sources because of their widespread popularity and the ease of data collection (Akcora, Carminati and Ferrari 2013, Celik and Dokuz 2018, Li, Goodchild and Xu 2013, Yuan, Jiang and Gidófalvi 2013). Yuan and Wang (2018) explored how data collection duration and sample size affect the effectiveness of using LBSM data to calculate two factivity space indicators. It showed that for the majority of users, their ROG and entropy (activity regularity/diversity) values grow as the data collection duration becomes longer. Lee et al. (2016) used 17-week long Twitter data to identify the differences between weekday and weekend activity spaces in southern Santa Barbara County, CA. Cheng et al. (2011) investigated 22 million Twitter check-ins across 220,000 users and found that users follow the “Levy Flight” mobility pattern and adopt periodic behaviors. Li et al. (2013) explored the spatial and temporal distributions of social media users in California and explored the socioeconomic characteristics of these users by using geo-referenced tweets collected from Twitter and Flickr.

In general, LBSM data can provide abundant individual activity records for little cost compared to traditional travel surveys. In addition, the location data collected from the built-in GPS of mobile phones provides a higher spatial accuracy. However, a series of potential data issues from LBSM data can also affect the effectiveness of analysis results when deriving human activity patterns from these datasets (Kaisler et al. 2013). Spatial data quality, such as low resolution, completeness, and consistency (Veregin 1999), plays a fundamental role in geographic analysis, therefore it is crucial to assess the effectiveness of LBSM data for human activity studies (Spielman 2014). In this study, we mainly focused on the data quality issues caused by the sparseness and low resolution of LBSM data in activity space stuides.

3.2.2 External Activity Space Indicators

Most of activity space studies differentiated the external and the internal characteristics of an activity space – the former measures its basic external morphology and the latter emphasizes the internal structure of an activity space. Various indicators have been developed to measure the external morphology of human activity space, such as the size, shape, and orientation of activity space.

Some studies showed that activity space can be represented as the minimum shape that includes all individual visited places within a minimum bounding geometry - minimum convex hulls (Fan and Khattak 2008, Lee et al. 2016). The minimum convex hull is a straightforward method to compute the region occupied by visited locations. However, they are not appropriate to use when the activity space is highly dispersed. An improved method called alpha shape method was developed to address the shortcomings of minimum convex hulls. It was firstly introduced in Edelsbrunner (1983) as a generalization of minimum convex hulls (Edelsbrunner, Kirkpatrick and Seidel 1983). Many examples proved that the alpha shape method represents the activity spaces more accurately than minimum convex hulls, and it was applied in many fields like pattern recognition, bio-informatics and sensor networks (Duckham et al. 2008, Fayed and Mouftah 2009).

In addition, historically, ellipse-based measures, such as the SDE and confidence ellipse, were also used to approximate activity spaces (Lefever 1926, Schönfelder and Axhausen 2003, Shannon and Spurlock 1976, Yuill 1971). Activity ellipses are generated based on distances and directions of an individual visited locations from the activity center (Gesler and Albert 2000).

Since ellipse-based representations were originally designed to exclude outlier locations, they are less affected by outliers than convex hulls. However, the main limitation of SDE is that it is an abstract representation of the area covered by a person rather than a full description of all the locations a person visited. Furthermore, the ROG has been widely used to represent the spatial dispersion and activity range of individuals' daily activities (Cheng et al. 2011, Song et al. 2010). As mentioned in Gonzalez et al. (2008), ROG is considered a robust indicator of activity scale and a measurement of the external morphology, which is less sensitive to outlier points.

3.2.3 Internal Activity Space Indicators

In addition to external characteristics, researchers also applied various measures to quantify the internal structure of individuals' activity spaces. Schönfelder and Axhausen (2003) calculated the shortest-path distance for each location visited by the travelers as an approximation of the actual paths based on individual origin-destination matrix and road network. Other network-based measures like standard travel time polygon and shortest-path spanning tree were applied to reveal the travel network structures of activity spaces (Sherman et al. 2005).

A branch of density-based measures is used to generate activity density surfaces for representing the intensity of activities in various spaces. Susilo and Kitamura (2005) and Kwan (2000) applied density-based surfaces to examine the spatial relationships between these density patterns by showing the spatial intensity of the locations. Schönfelder and Axhausen (2003) mapped the weighted activity space to show the distribution of frequently visited locations using kernel density estimation. Li, Li and Shan (2017) found that higher tweets' density areas are surrounded by commercial and institutional places.

Moreover, entropy is often used to indicate the randomness of activity patterns, which is invaluable in determining the likelihood of users returning to previously-visited locations and predicting future trips. It is used to show the movement among the most frequently visited locations (i.e., POIs) and quantify the probabilistic distribution of visiting different locations. It proved to be less impacted by outlier points (Song et al. 2010, Yuan and Raubal 2012, Yuan and Wang 2018). Similarly, Gong et al. (2016) inferred trip purposes by analyzing their visit probability based on POIs and drop-off points distribution and uncovered travel patterns from taxi trajectory data. Some studies identified activities by combining human activities with POIs. Xie, Deng and Zhou (2009) assigned the POI type as the activity purpose of a person. Huang, Li and Yue (2010) extracted a person's potential activity sub-trajectories from their entire travel route by defining the spatiotemporal attractiveness of POIs. Phithakkitnukoon et al. (2010) found a strong correlation of activity patterns within people who work in the same area.

3.3 Methodology

As discussed in the introduction chapter, this study evaluates the effectiveness of measuring intra-individual activity spaces indicators from LBSM data.

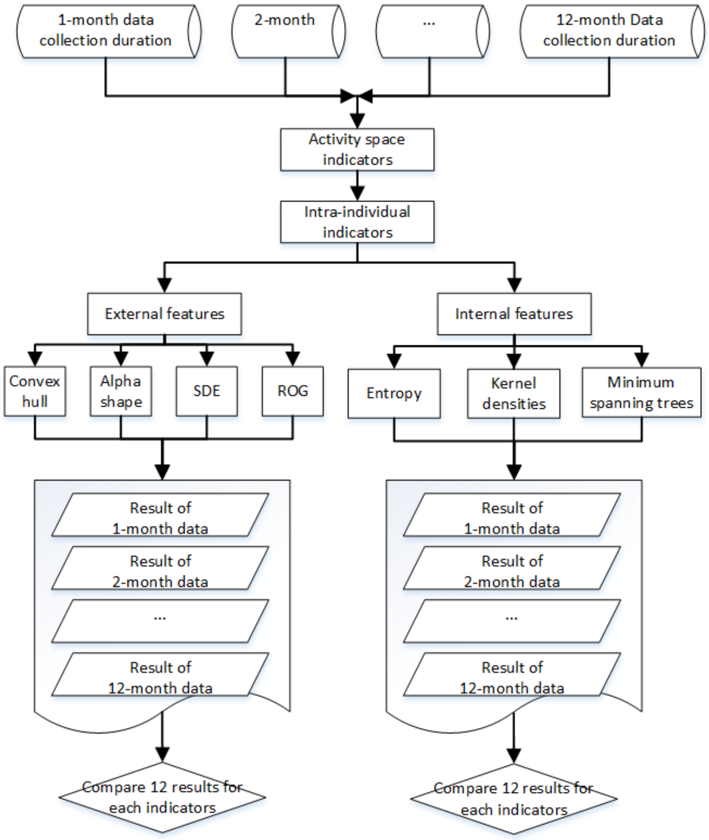


Figure 3.1 Workflow of Effectiveness Analysis on Intra-individual Indicators.

We chose four external indicators to calculate the LBSM user's activity space size and three internal indicators to represent activity space structures and regularity among various kinds of intra-individual activity space indicators (see Figure 3.1).

3.3.1 Define Activity space indicators

Generally, human activity space can be interpreted by two aspects, external morphology and internal structures (e.g., regularly visited locations, network structure, etc.). We chose the most frequently used indicators from previous activity studies (Lee et al. 2016, Schönfelder and Axhausen 2004, Golledge and Stimson 1997, Schönfelder and Axhausen 2002, Song et al. 2010, Sherman et al. 2005, Yuan et al. 2012, Kwan 2000).

(1) External activity space indicators

- Minimum convex hull: In mathematics, the minimum convex geometry is defined as a polygon that contains all points and has no internal angles greater than 180 degrees on a 2-dimensional plane (Fan and Khattak 2008, Lee et al. 2016, Andrew 1979). In this research, it shows a unique activity space which contains all check-in locations from an LBSM user (Figure 3.2). The minimum convex hull is straightforward to compute, but it is imperfect if the person’s activity space is irregularly shaped, because the check-in outliers make activity space inaccurate.

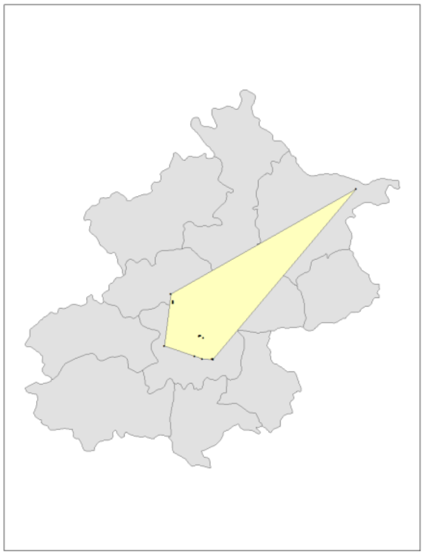


Figure 3.2 A Sample of Minimum Convex Hull.

- Alpha shape: The alpha-shape associated with a set of points is a generalization of the minimum convex hull. It is a family of piecewise linear simple curves in the Euclidean plane associated with the shape of a finite set of points (Edelsbrunner et al. 1983, Akkiraju et al. 1995, Barbosa, da Fontoura Costa and de Sousa Bernardes 2003, Duckham et al. 2008). However, unlike the minimum convex hull, alpha shape is constructed as a non-convex enclosure on a set of points (Figure 3.3). Hence, it provides a more accurate boundary than the convex hull and is less affected by outliers (Fayed and Mouftah 2009).

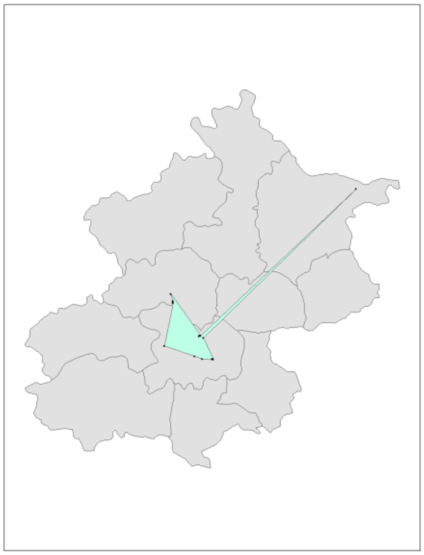


Figure 3.3 A Sample of Alpha Shape.

- Standard Deviational Ellipse (SDE): this indicator calculates the standard deviation of *x* coordinates and *y* coordinates from the mean center of an individual's check-ins to define the axes of the ellipse (Lefever 1926, Shannon and Spurlock 1976, Yuill 1971, Schönfelder and Axhausen 2003). The ellipse allows us to see the shape and the orientation of a user's activity space. In this case, this study calculates the standard distance separately in the *x* and *y* dimensions by measuring the trend for a set of check-ins of an LBSM user. These two measures define the major and minor axes of an ellipse. This ellipse is referred to as the standard deviational ellipse, since the method calculates the standard deviation of the *x* coordinates and *y* coordinates from the mean center to define the axes of the ellipse (Figure 3.4).

$$SDEx = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n}} \tag{1}$$

$$SDEy = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{Y})^2}{n}} \tag{2}$$

where x_i and y_i are the coordinates for check-in, $\bar{X}$ and $\bar{Y}$ represents the mean center for the check-ins of one LBSM user, and *n* is the total number of check-ins from this user.

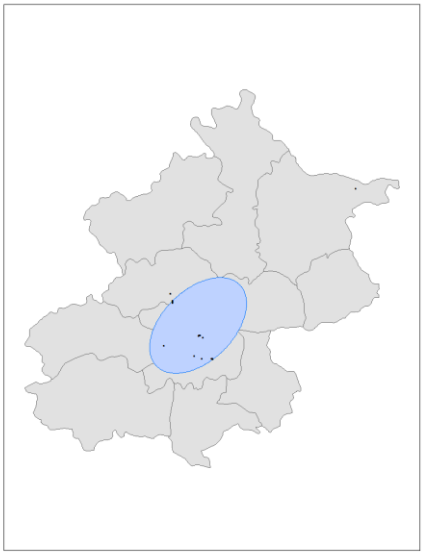


Figure 3.4 A Sample of SDE.

- Radius of Gyration (ROG): it is considered an indicator of activity scale and a measurement of the external morphology. ROG represents the activity range of individual activity space around their center check-in footprint (Cheng et al. 2011, Song et al. 2010, Xu et al. 2015). Compared to SDE, it only calculates one parameter (i.e., radius) to represent the scale of activity spaces (Figure 3.5). ROG has been widely used to represent the spatial dispersion of individual daily activities. It is defined as:

$$\mathrm{ROG} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\vec{r_i} - \overrightarrow{r_m}\right)^2} \tag{3}$$

$$\mathrm{Area}_{ROG} = \pi * \mathrm{ROG}^2 \tag{4}$$

where *n* refers to the number of check-in locations of a given user; *r_i* is the geographical coordinate of each check-in location; *r_m* refers to the centroid of all check-in points of a given user.

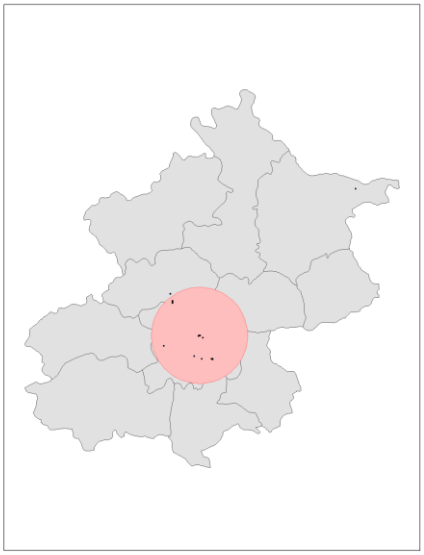


Figure 3.5 A Sample of ROG.

According to the definition of ROG, there is an advantage of ROG over the other three intra-indicators. Only the ROG indicator can capture a user's activity space when the check-in locations are located along a straight line (Figure 3.6). In this case, although the user moved in different locations, there is no 2-dimensional activity space formed. Convex hull, alpha shape, and SDE methods are not able to detect the 2-dimensional activity space of this user. However, ROG calculates the centroid of the check-in locations and a mean distance among all check-ins to the centroid location, therefore the activity space of one user is represented within the distance (i.e. ROG value) from the centroid location.

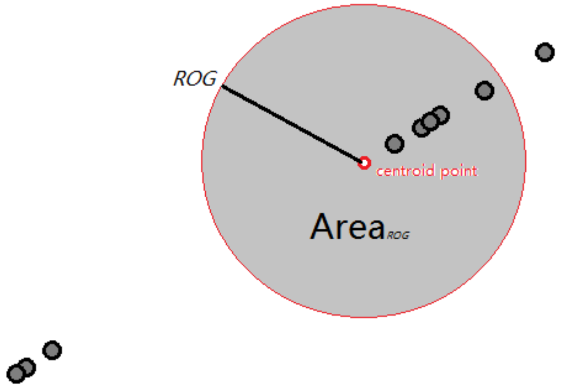


Figure 3.6 A Line-Shaped Check-in Distribution.

(2) Internal activity space indicators

- Entropy: it is often used to indicate the randomness of activity patterns, which is invaluable for determining the likelihood of users returning to previously-visited locations and predicting future trips (Song et al. 2010, Yuan and Raubal 2012, Yuan and Wang 2018). Entropy is defined as:

$$\mathrm{E} = -\sum_{i=1}^{N} p_i \log_2 p_i \tag{5}$$

where p_i refers to the probability of a given user checking in at the same place i, and N stands for the total number of places where this user checked in. It is considered an indicator of the internal structure and randomness of activity spaces.

As can be seen from Figure 3.7, user A visited more disperse locations in the neighborhood while user B preferred to visit particular locations instead. It can be considered that activity diversity of user A is higher than B. Entropy indicator is used to quantify the degree of the regularity of an individual. Higher entropy value represents higher randomness of a pattern (e.g., the entropy of user A is higher than user B).

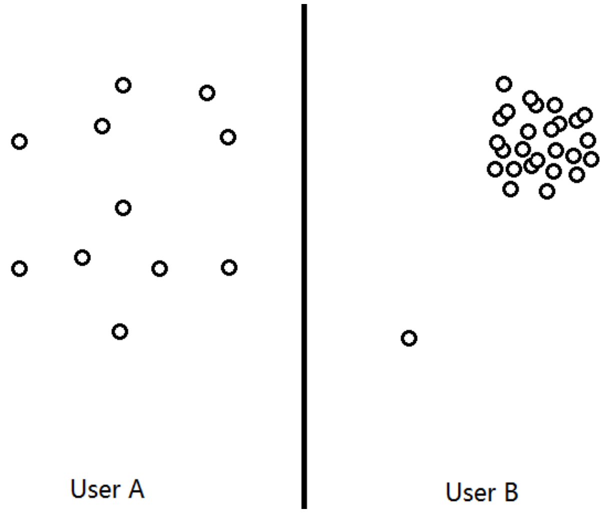


Figure 3.7 A Sample of Different Entropy Patterns.

- Minimum Spanning Trees (MST): it reveals the movement network among regularly visited locations (Schönfelder and Axhausen 2002, Sherman et al. 2005, Gower and Ross 1969). An MST is a subset of the edges of a connected, (un)directed graph that connects all the locations together without any cycles. It is a spanning tree whose sum of edge weights is the smallest. In this study, we calculate all possible connections among all check-in locations for each user and select the minimum route (black lines) as the MST distance (Figure 3.8). This indicator shows the minimum connections distance of all the check-in locations. It generates the spanning tree whose sum of connection cost is the smallest. This indicator quantifies the connection complexity of an LBSM user's visiting locations in his/her activity space based on the structure among his/her check-in locations.

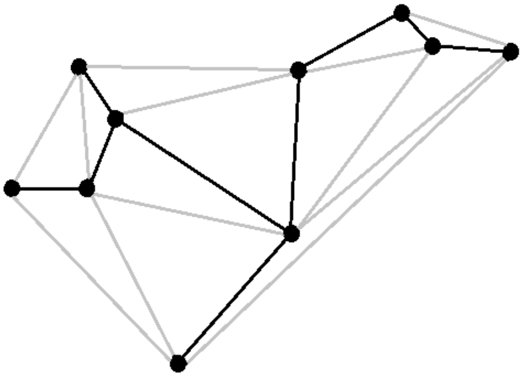


Figure 3.8 A Sample of Minimum Spanning Trees.

- Kernel density: it shows the density distribution of activity locations. The basic process of a kernel density analysis is to transform a point pattern (such as the set of activity locations visited) into a continuous representation of density in space (Kwan 2000, Schönfelder and Axhausen 2003, Susilo and Kitamura 2005). The more check-ins in an area means the more often the neighborhood is visited. The highest density of activity location occurs in the most frequently visited areas and the lowest exists within the least visited ones. In this case, we calculate the location with the highest check-in density between 8 pm to 5 am (i.e., possible home location) for each LBSM user to see how this location changes with an increasing data collection duration. Figure 3.9 shows a sample of highest density (red points) displacement by accumulating data amount.

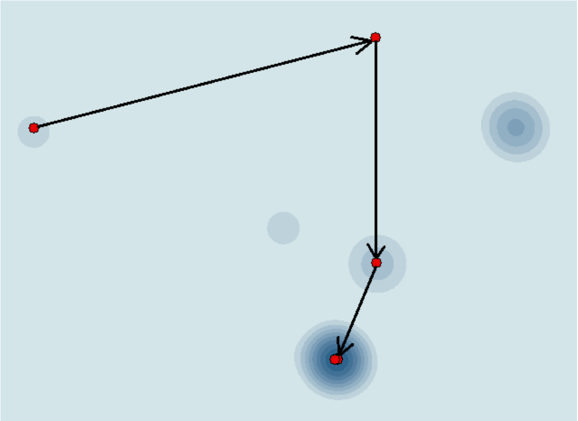


Figure 3.9 Displacement of the Highest Density Check-in Location.

3.3.2 Test Effectiveness of Activity Space on Different Indicators and Data Sizes

This chapter tested the effectiveness of LBSM data in modeling intra-individual activity space. We calculated four external activity space indicators to represent the size and shape of activity spaces as defined in 3.3.1, which are convex hull, alpha shape, SDE, and the ROG. We compared the results of these four different indicators in modeling LBSM activity spaces. Besides the external indicators, this study also employed three internal indicators to model activity spaces: entropy to show the probabilistic distribution of visiting different POIs; minimum spanning trees to reveal the network among visited locations, and kernel densities to show the most checked-in locations. These indicators were used to represent the internal structure of activity spaces and the regularity of visited locations.

This study also tested how data quantity affects the magnitude of activity space indicators. We investigated how an individual activity space changes when applying different data collection durations to activity space indicators. In this study, "data quantity" and "data collection duration" were used interchangeably. We chose data collection duration (e.g., 1 month, 2 months) instead of the exact number of check-ins (e.g., 1,000 records, 2,000 records) for two reasons: (1) to be consistent with other social media analysis, as most data collection campaigns are conducted based on a chronological circle (e.g., weeks or months); and (2) to collect our user data under the same study period so that they are comparable. In other words, we tested how LBSM data collection duration affects the computation of activity space indicators. In details, we applied different LBSM data sizes to the seven intra-individual indicators respectively.

First, we calculated the magnitude of 7 activity space indicators based on 12 different data collection durations – from 1 month to 12 months. Then, we used a linear regression model with logarithmic transformation to approximate how the magnitude of each indicator changes with different data collection durations. We are interested in exploring whether an indicator approaches a steady point as the number of months increases, and if so, how to approximate this limit using a mathematical model. For example, the indicator may increase, decrease, or fluctuate as the number of data increases. At last, we conducted a model fitting to approximate the limit of activity space indicators in Beijing, Shanghai, and Guangzhou. Understanding how the indicators change can provide useful insights for choosing an appropriate data collection duration in future studies.

3.4 Results

3.4.1 Results of Activity Space Indicators

We calculated four external and three internal indicators for each Weibo user using the data collected from April 2015 to March 2016. As mentioned in Chapter 2, to ensure that we have adequate information to extract individual users' activity patterns, we only considered users with at least 10 check-ins during the study time span from April 2015 to

March 2016. After data pre-processing, we obtained 6,215 users over three cities. The summary of the data is listed in Table 3.1.

Table 3.1 Data Summary for Indicator Calculation.

April 2016-March 2016	Amount of check-ins	Amount of users
Beijing	50,355	2,800
Shanghai	33,122	1,776
Guangzhou	30,608	1,639

After converting the longitude and latitude coordinates of check-in locations to a planar coordinate system (i.e., Pseudo-Mercator -- Spherical Mercator), we compared four external activity space indicators for each Weibo user by grouping their check-ins through their unique user IDs. We used the average value (Table 3.1 and Table 3.2) for each indicator to see how the scale of activity spaces was different based on different indicators (Figure 3.10).

Table 3.2 Average Value of 4 External Indicators.

Average	ROG area(km^2)	SDE(km^2)	Convex hull(km^2)	Alpha shape(km^2)
Beijing	295.0612	186.7965	125.7324	115.0579
Shanghai	151.3380	81.1435	70.3178	64.0062
Guangzhou	112.4128	48.7173	46.8779	41.6438

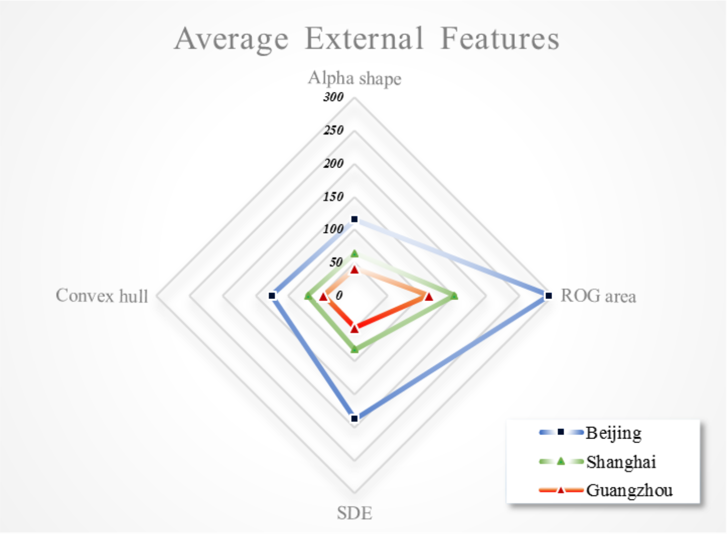
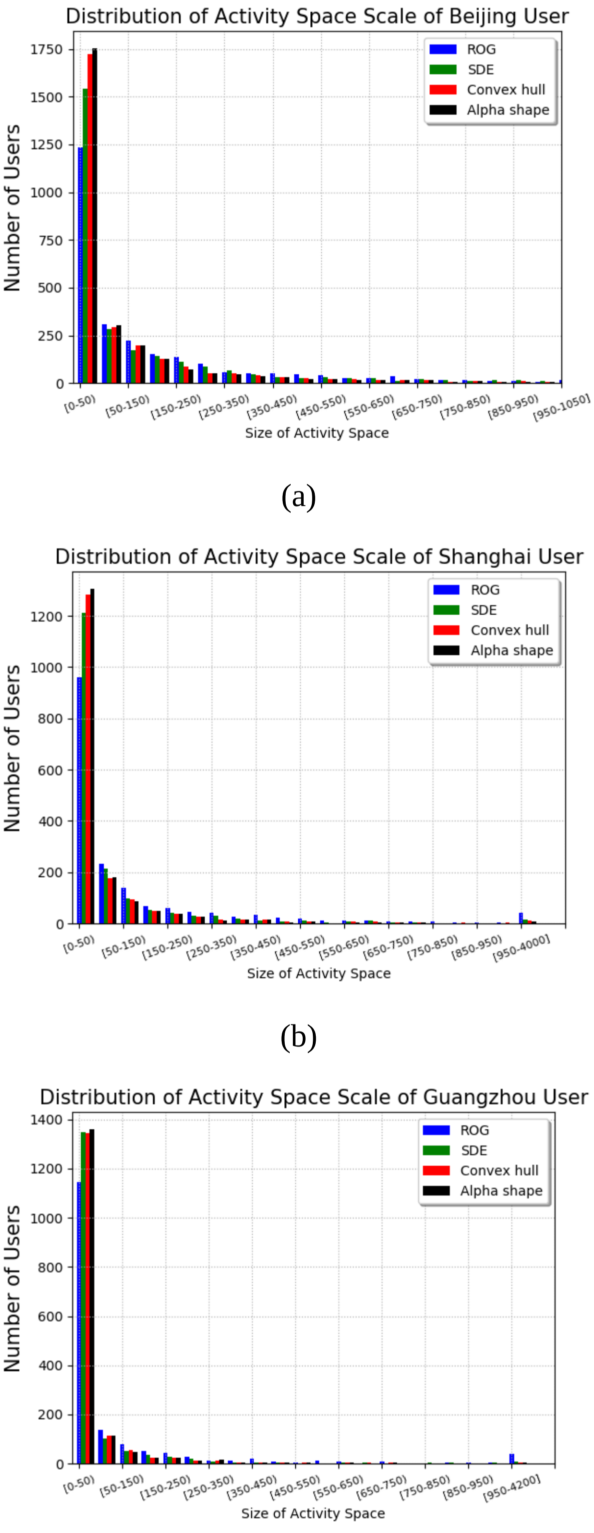


Figure 3.10 Scale of Activity Space based on Different External Indicators.

Figure 3.10 shows a radar chart of activity space of three study areas. Each axis represents one of the external activity space indicators, which are convex hull, alpha shape, SDE and ROG area. The radii range from 0 km^2 to 300 km^2 and the data points show the scale of the magnitude of the activity space. The overall magnitude of the average activity space of Beijing users is the largest among all three cities, followed by Shanghai users, and Guangzhou users have the smallest average activity space. It shows that an individual's activity space is possibly related to the size of the city. In this study, the bigger the city, the larger the individual average activity space is.

While comparing the performance of four external activity space indicators, it is clear that the average activity spaces represented by ROG areas are larger than the other three indicators in all three study areas. The possible reason is that the ROG method is less sensitive to outliers and covers the most possible activity spaces of the LBSM user. The areas from convex hull and SDE are similar for Shanghai and Guangzhou, but not for Beijing. The average activity space calculated by SDE is 61 km^2 larger than convex hull in Beijing, 11 km^2 larger in Shanghai, and 2 km^2 larger in Guangzhou. From the tabular data in Table 3.2, it seems the bigger the city, the larger the difference between the convex hull and SDE indicators in this case. It also shows that alpha shape is less capable to differentiate the three cities. The differences of activity spaces represented by convex hull and alpha shape are the least among all the indicators. Moreover, the average space sizes represented by alpha shape are slightly smaller than the convex hull in all three study areas. It proves that the alpha shape indicator describes a more accurate boundary and is less affected by outliers.



(a)

(b)

(c)

Figure 3.11 Distribution of 4 External Indicators: (a) Beijing; (b) Shanghai; and (c) Guangzhou.

Besides showing the average values of four indicators in three cities, we also plotted the histograms to show the distribution of each indicator. As can be seen from Figure 3.11, all four indicators show power law distribution with high values on the left-hand side and a long tail of outliers on the right-hand side of the distribution. Activity space sizes of 80% LBSM users are within 50 km^2 in all three cities. The results also follow the “80/20” rule (Jiang, Yin and Zhao 2009) where the majority of Weibo users (more than 80%) have their activity space sizes within 50 km^2 and only a few users (less than 20%) have extremely large activity space sizes. Activity spaces with more outliers (e.g., check-ins far away from the most frequently visited locations) are more sensitive to the different activity space indicators because their activity scales are highly affected by the outliers.

In addition, we are interested in the disparity of individuals’ activity spaces represented by different indicators. We selected the ROG area as a base indicator to evaluate the absolute differences between ROG and the other three external activity space indicators. Figure 3.12 shows the absolute value of the difference between the base indicator (i.e., ROG) and the other three indicators respectively.

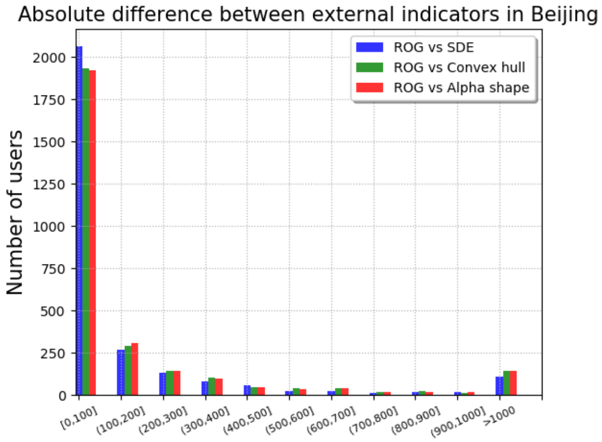


Figure 3.12 The Differences among Four External Indicators in Beijing.

As presented in Figure 3.12, the absolute difference between each pair of the external indicators follows a power law distribution. The number of users whose activity space differences were less than 100 km^2 occupied almost 70% of all the users in Beijing. It shows that the activity space of the majority of users did not change much although presented by different indicators, and only a few users have huge differences in activity space scales. Specifically, only less than 5% of the Beijing users show substantial differences (i.e. absolute difference greater than 1,000 km^2) when applying different activity space indicators.

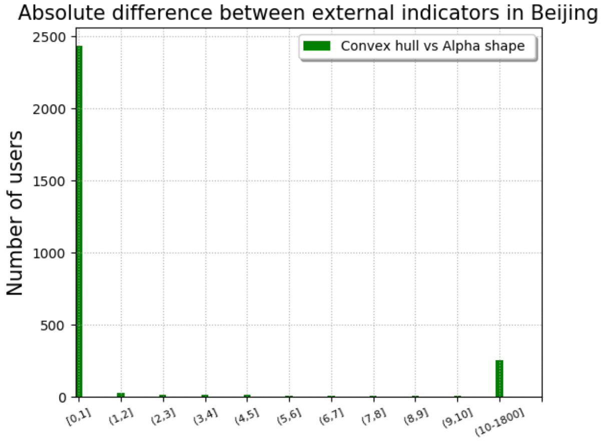
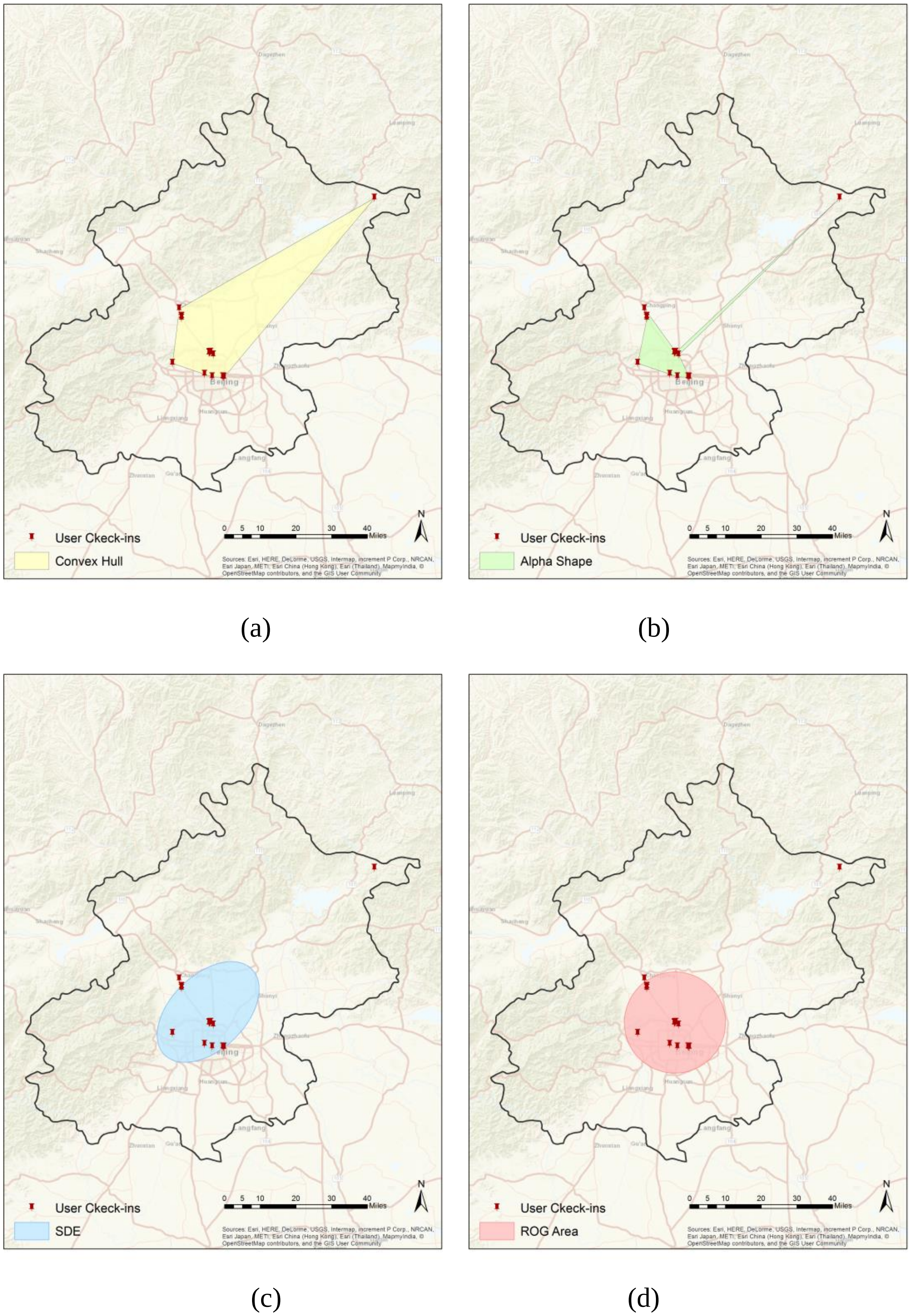


Figure 3.13. Comparison of Convex Hull and Alpha Shape.

As can be seen from Figure 3.13, the difference between convex hull and alpha shape is even smaller. 45% of Beijing users' activity space size differences are within 0.1 km^2. As can be seen from Figure 3.14(a)-(b), alpha shape is less affected by outliers and can measure activity space more accurately than convex hull. In summary, alpha shape, as an improved method of the convex hull, resulted in a more accurate and smaller activity space than the convex hull. We also chose an example to show why some users' activity space sizes changed dramatically with different indicators (Figure 3.14 and Table 3.3).



Sources: Esri, HERE, DeLorme, USGS, Intermap, increment P Corp., NRCAN, Esri Japan, METI, Esri China (Hong Kong), Esri (Thailand), MapmyIndia, © OpenStreetMap contributors, and the GIS User Community

(a) (b)

(c) (d)

Figure 3.14 A Sample to Illustrate Activity Space Difference by Different Indicators: (a) convex hull; (b) alpha shape; (c) SDE; and (d) ROG area.

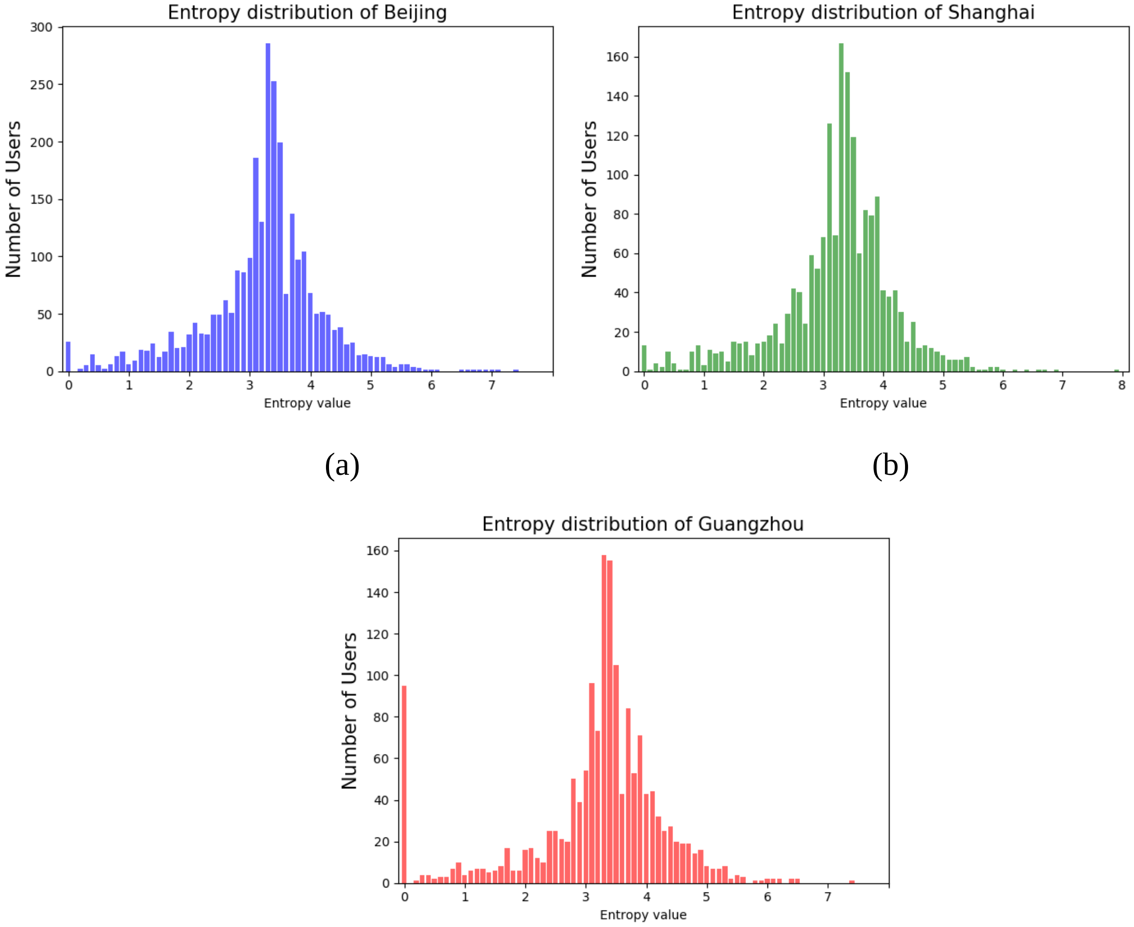
Table 3.3 Activity Space Sizes of the User.

Indicators(km^2) User ID	ROG area	SDE	Convex hull	Alpha shape
268874****	1,632	1,457	2,170	400

As indicated by Figure 3.14, most of the places this user visited are located in central Beijing, however, there is one located far away from the others. The result shows that the activity space represented by alpha shape is the smallest (see Figure 3.14(b) and Table 3.3). In this case, different from the convex hull, alpha shape accurately reduced the biases affected by the outlier point located in northeast Beijing; however, SDE is also highly affected by the outlier point. Figure 3.14(c) shows that the direction of the long axis of the SDE is toward the outlier in northeast Beijing while most of the check-ins are not in that direction. Figure 3.14 (d) shows that the ROG area can approximate an activity space that is not sensitive to the outlier.

We also calculated three internal activity space indicators for Weibo users to describe the travel diversity, network structures among those check-in locations, and the movement pattern within the activity space. We plotted the distribution for each indicator in three study areas.

Figure 3.15 shows the histograms of entropy values in three cities. Our analysis shows similar distributions for three cities, with the mean entropy value of 3.2546 for Beijing, 3.2852 for Shanghai, and 3.2136 for Guangzhou. The results revealed that the average values of entropy seem not to vary across three cities, indicating that the size and structure of the cities have little impact on the randomness of an individual’s activity space. There is a notable pattern of entropy values in Guangzhou that is different than the patterns in Beijing or Shanghai: As can be seen from Figure 3.15(c), a total of 95 users have their entropy value equal to 0, which means that these users posted from the same location during the entire data collection duration. It suggests that 1) these users stayed in the same place when they posted, (2) these users are more likely to be residents rather than visitors in Guangzhou, (3) their activity space sizes were calculated as zero because there was no movement detected.



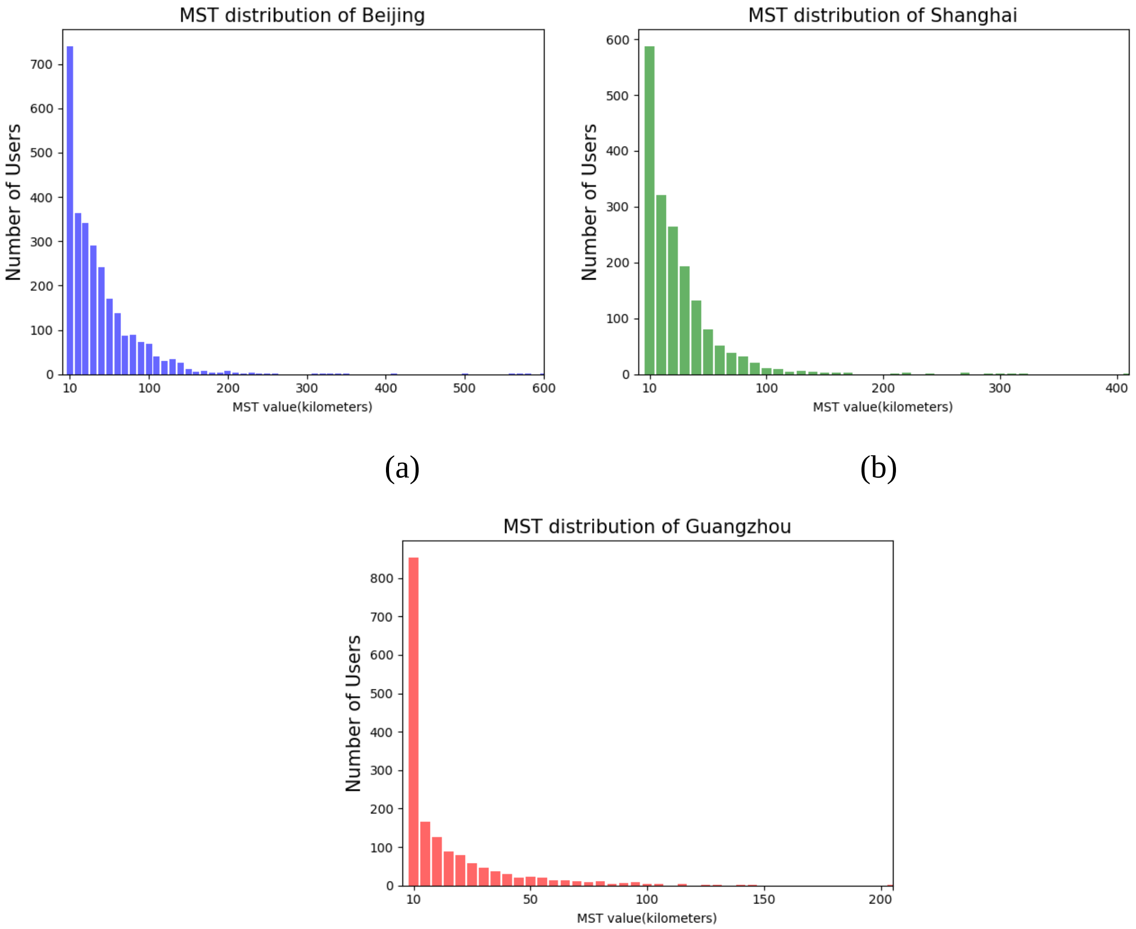
(a)

(b)

(c)

Figure 3.15 The Distribution of the Number of Users and Entropy Values: (a) Entropy distribution of Beijing; (b) Entropy distribution of Shanghai; and (c) Entropy distribution of Guangzhou.

For the MST, our analysis results are shown in Figure 3.16.



(a) (b)

(c)

Figure 3.16 The Distribution of the Number of Users and MST Distances: (a) Beijing; (b) Shanghai; and (c) Guangzhou.

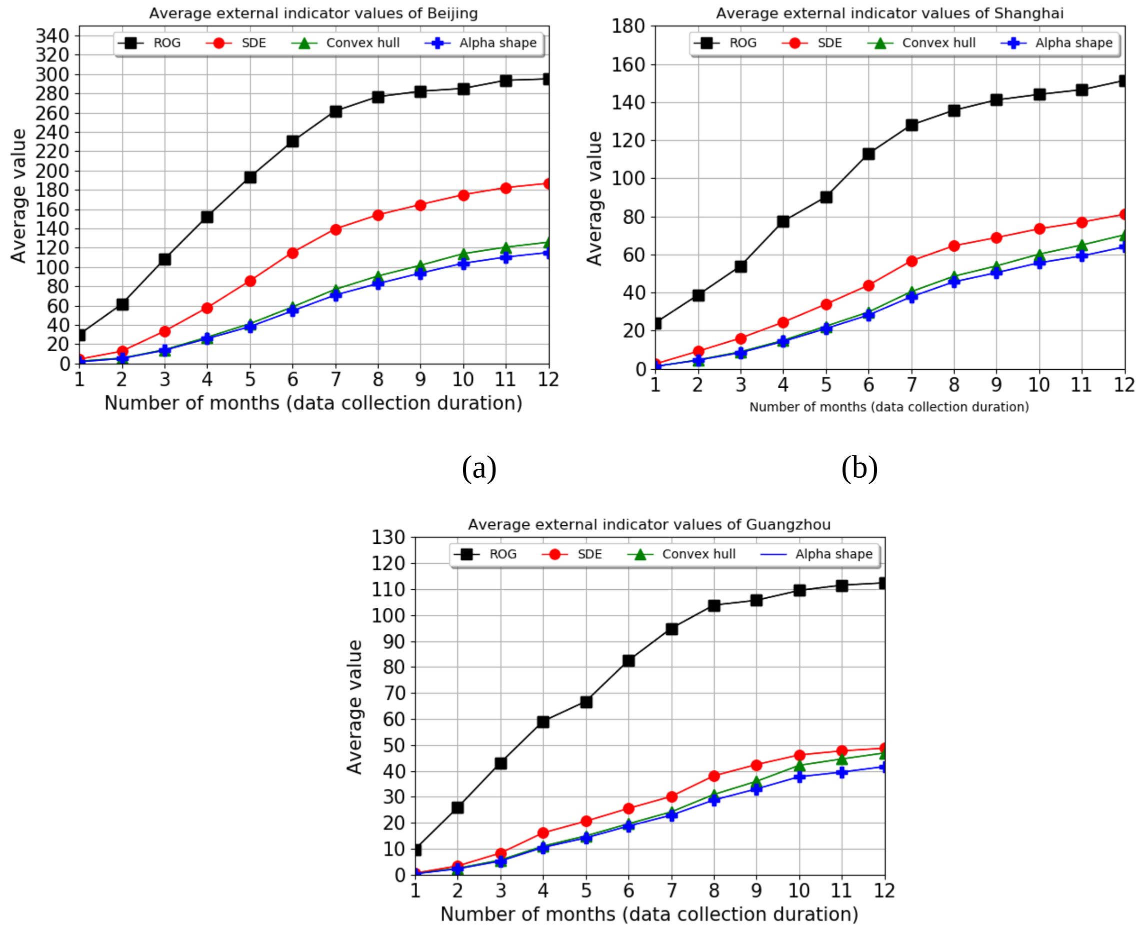
We calculated and compared the MST distance for the users in three cities. Figure 3.16 shows that majority of LBSM users have a short network connecting distance among their check-in locations in all three cities. In brief, the MST method connects all the check-in locations together using the shortest path. The MST distance describes the shortest connection path among the check-in locations. The further the MST distance, the more connection complexity among user visited locations. It showed that a large proportion of LBSM users tend to use a small set of locations for their daily activities. The average MST distance of Beijing is the longest for 41.58 km, and then of Shanghai for 28.06 km, while Guangzhou is the least for 18.03 km. The results of the MST indicator demonstrated that the average MST distance of Beijing users is much further than that of Shanghai and Guangzhou. This is potentially determined by the city's size, planning, and infrastructure. From the MST indicator, we can infer that Beijing has a more complex environment and a larger activity space extent than Shanghai or Guangzhou.

The analysis of kernel density was conducted by calculating the movement of locations with the highest check-in density between 8 pm to 5 am (i.e., possible home location) for each LBSM user. We tested how this most visited location changes with an increasing data collection duration in chapter 3.4.2.

3.4.2 Results of Activity Space on Different LBSM Data Sizes

As an effectiveness test to explore the correlation between activity space indicators and the amount of data utilized, we calculated the average values for all intra-individual indicators based on different data collection durations (from 1 month's data up to 12 months' data) in three cities.

(1) External activity space indicators



(a) (b)

(c)

Figure 3.17 Scale of Activity Space based on Different Collection Durations

(a) Beijing; (b) Shanghai; (c) Guangzhou.

Figure 3.17 shows how convex hull, alpha shape, SDE and ROG values change with the different amount of data used. As can be seen, in all three cities, all four indicators show an increasing trend with a longer data collection period; however, the increasing trend slows down and the indicator approaches a limit value as the amount of data continues to grow. To further quantify the change of activity space indicators with different data collection durations, we plotted the percentage of increase (p_i) of the indicator (Figure 3.18), defined as:

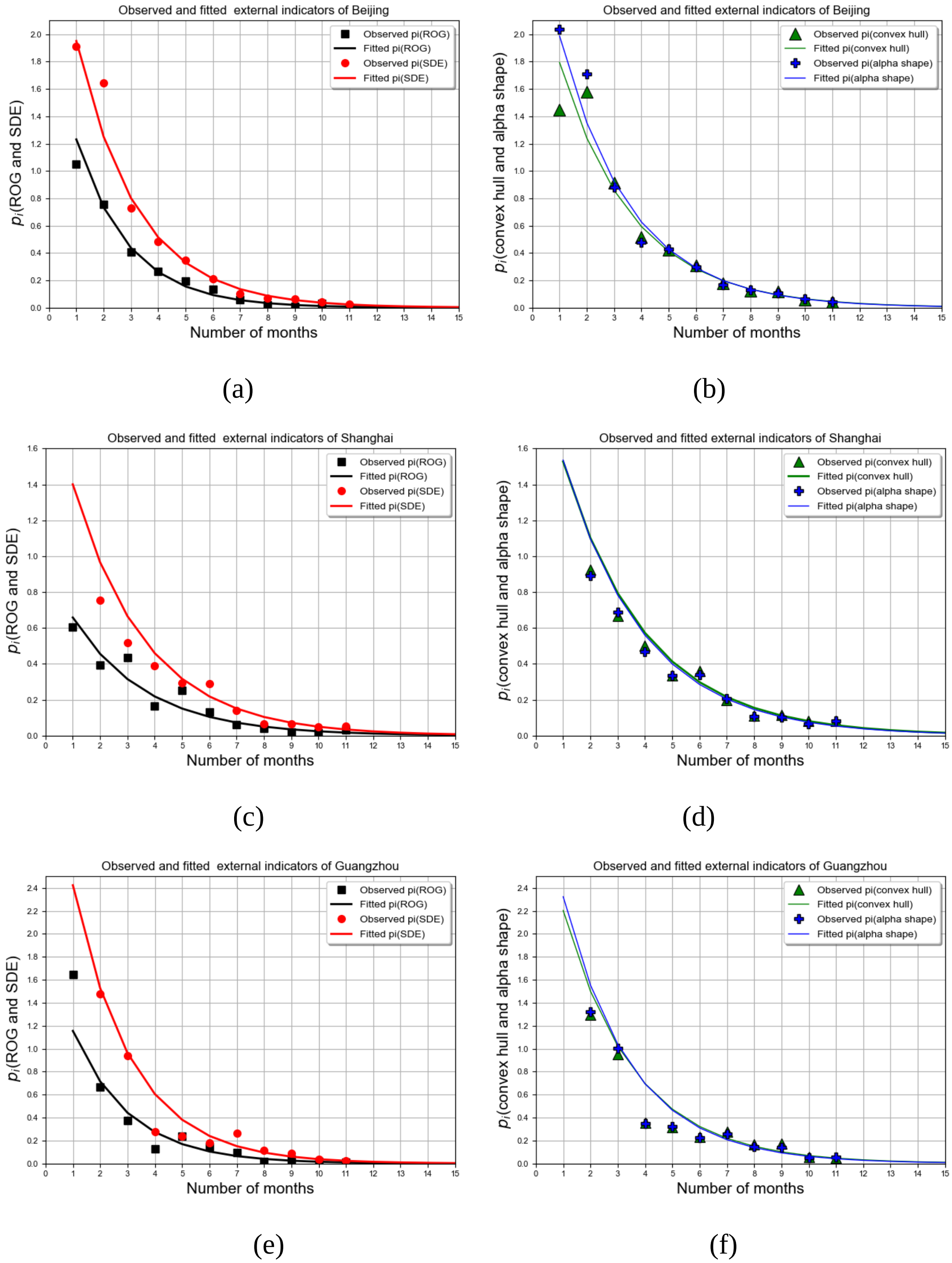
$$p_i = \frac{x_{i+1} - x_i}{x_i} \tag{6}$$

where x_i stands for the value of indicator x calculated using i months' worth of data. As shown in Figure 3.18,the increase rate of all indicator values in three cities decreases when the amount of data increases. Since the correlation does not appear to be linear, we apply a logarithmic transformation on pi and construct the following regression model:

$$\log(p_i) = am + b \tag{7}$$

where m is the number of months of data used in the analysis, and a and b are the coefficient and intercept of the fitted regression model.

Figure 3.18 shows that the increasing trend slows down and approaches zero as the data collection duration increases to 12 months. It indicates that these indicators reach a limit value as the amount of data continues to grow. When the data collection duration reaches 9 months, the increasing rates of all indicators are less than 0.1. This result is consistent with the assumption of time geography (Hägerstrand, 1970), where an individual's daily activity space is restricted to a certain spatial range due to physical constraints (e.g., moving speed), administrative boundaries, lifestyles, and so on.



(a) (b)

(c) (d)

(e) (f)

Figure 3.18 Increasing Rate of 4 Indicators based on Different Collection Durations: (a) ROG vs. SDE in Beijing; (b) Convex hull vs. alpha shape in Beijing; (c) ROG vs. SDE in Shanghai; (d) Convex hull vs. alpha shape in Shanghai; (e) ROG vs. SDE in Guangzhou; and (f) Convex hull vs. alpha shape in Guangzhou.

Table 3.4 Comparison of Observed Indicators and the Simulated Limit (Beijing).

Beijing	average value	simulated limit value	value/simulated (%)
ROG	295.0612	298.083	98.99%
SDE	186.7965	194.5516	96.01%
Convex hull	125.7324	138.9209	90.51%
Alpha shape	115.0579	125.8575	91.42%

Table 3.5 Comparison of Observed Indicators and the Simulated Limit (Shanghai).

Shanghai	average value	simulated limit value	value/simulated (%)
ROG	151.3381	157.0538	96.36%
SDE	81.14359	87.48515	92.75%
Convex hull	70.31781	81.55756	86.22%
Alpha shape	64.00625	74.23715	86.22%

Table 3.6 Comparison of Observed Indicators and the Simulated Limit (Guangzhou).

Guangzhou	average value	simulated limit value	value/simulated (%)
ROG	112.4129	114.1602	98.47%
SDE	48.71733	50.72552	96.04%
Convex hull	46.87795	51.81313	90.48%
Alpha shape	41.64381	45.24414	92.04%

As can be seen from Tables 3.4-3.6, when using 12-month average data, the calculated data is very close to the approximated limit value. ROG has the largest at 96% of the observed value in all three cities. This shows that ROG as an indicator for describing external activity space is not sensitive to the outlier. This result can be interpreted from multiple perspectives.

On the one hand, the simulated ROG and other indicators' limits provide quantitative evidence to interpret users' activity scale in a certain city. For example, in Beijing, the average ROG area is approximately 295 km^2, which is larger than the limits in Shanghai (151 km^2) or Guangzhou (112 km^2). This is potentially determined by the city's boundary, planning, and structure. The same analysis can be extended to different cities to study the impact of urban setting on activity space. On the other hand, we further confirmed that one-year's data is capable of capturing at least over 86% of the variability of all external indicators in three study areas. Future studies can adopt a similar methodology to determine a balance point between data quantity and analytical precision. In addition, the differences among ROG, SDE, convex hull and alpha shape are worth noting. For example, with 12 months' worth of data in Shanghai (Table 3.4), the calculated ROG is able to reflect over 96% of the limit value; however, for convex hull and alpha shape, this proportion drops to 86%, suggesting that various indicators may have a different level of sensitivity toward data quantity. Similar patterns exist for Shanghai and Guangzhou, indicating that convex hull and alpha shape values require a longer data collection period to stabilize.

(2) Internal activity space indicators

The previous chapter addressed the effectiveness of external indicators and data sizes in modeling activity space. Here, we illustrate the effectiveness of indicators in modeling the internal structure of activity spaces.

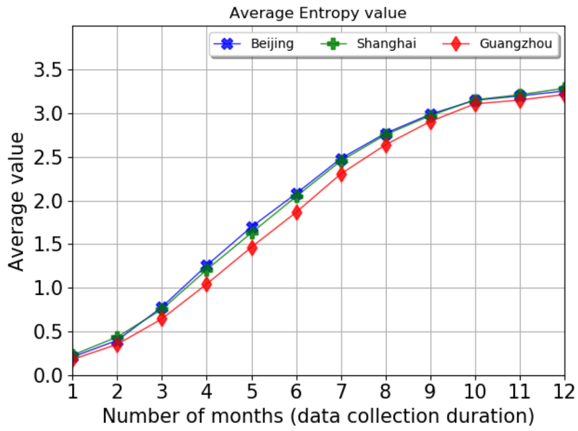


Figure 3.19 The Average Entropy Value of Users in Three Study Areas.

As can be seen from Figure 3.19, it follows the pattern that the average entropy values increase as the amount of data increases. The increasing trend slows down and approaches zero as the data collection duration increases to 12 months. It indicates that the average entropy value reaches a limit as the amount of data continues to grow in all three study areas. However, different from the external activity space indicators, our result shows similar distributions of entropy in three study areas, indicating that the size and structure of the cities have little impact on the randomness of individuals' activity spaces.

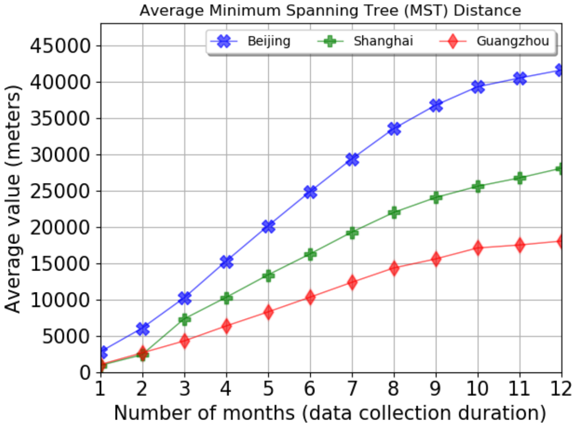


Figure 3.20 The Average MST Distance of LBSM users in Beijing, Shanghai, and Guangzhou.

Figure 3.20 shows how the MST distances change with the different amount of data used in three study areas. It also follows the same pattern that the average MST values increase as the amount of data increases. However, the increasing trend slows down and the indicator approaches a limit value as the amount of data continues to grow. The average MST distances of users in three cities are very different from each other. The average MST distance of Beijing users is 23 km larger than Guangzhou users. Although the MST indicator quantifies the shortest connection distance of check-in locations in activity space, it is also limited by the scale of a city as well as the size of the activity space.

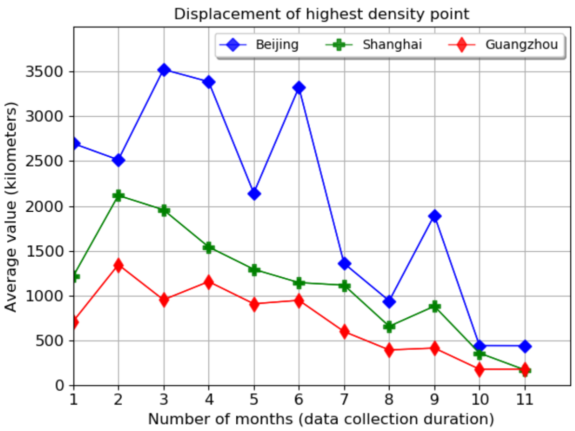


Figure 3.21 Average Displacement of the Highest Density Check-in Location.

In addition, we defined the most frequently visited location place as the highest density location of a given LBSM user. We calculated the highest density locations for each user by creating the kernel density surface of each user. We selected check-in points between 8 pm to 5 am to estimate the possible home location of LBSM users. Similar to external indicators, we applied 12 different data sizes to see how the most frequently visited location (i.e., the potental "home location") changes with different data collection durations under the same density surface prameter settings. Figure 3.20 shows that as the amount of data increases, the average displacement of the estimated home location in all three study areas decreases. In other words, when the data collection duration increases, the highest density location tends to stablize. Figure 3.22 shows an example of the displacement of the highest density location (red points) with different amounts of data.

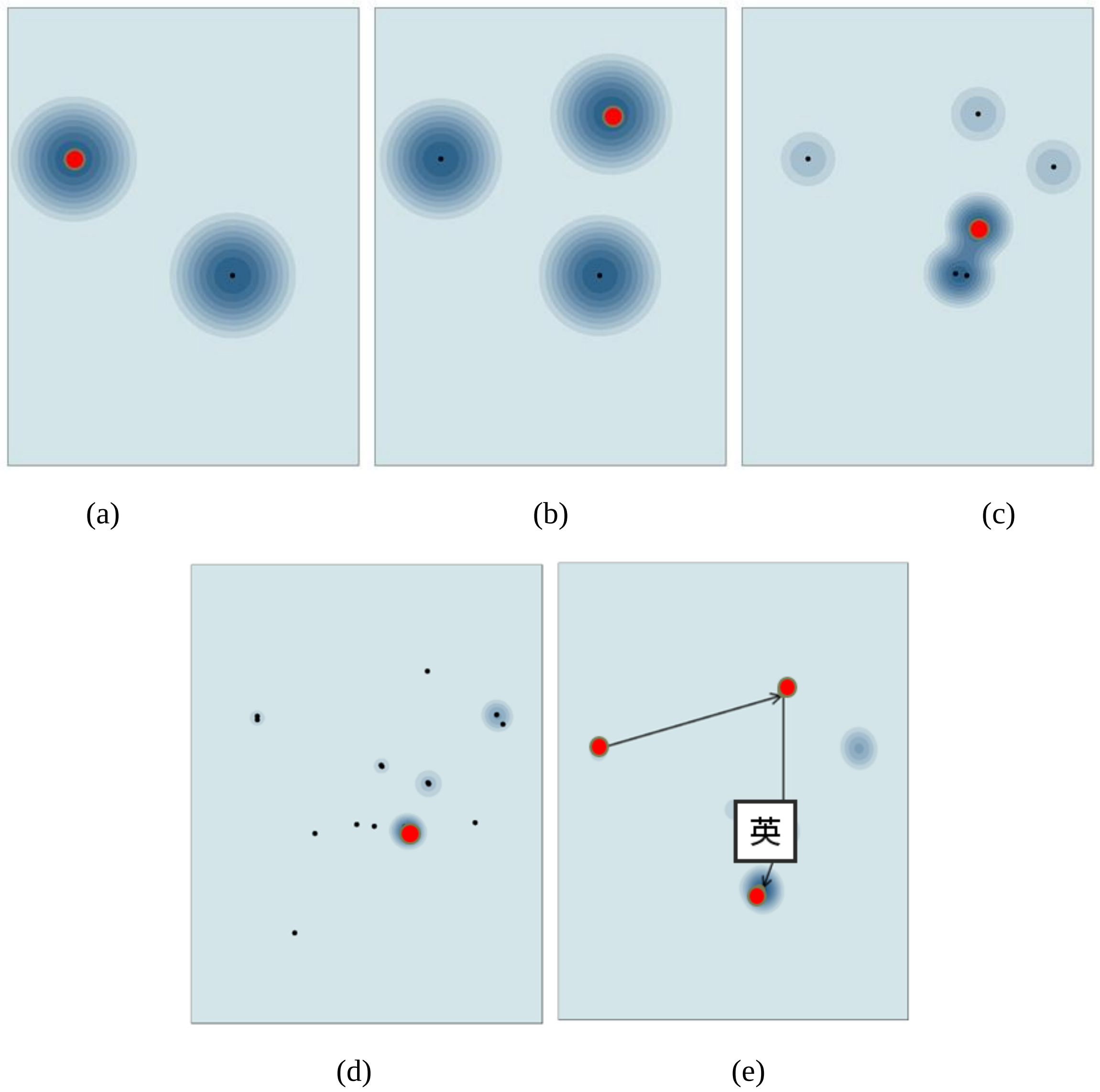


Figure 3.22 Analysis Procedure of Displacement of One Weibo User:

(a) 1 month's data; (b) 2 months' data; (c) 3 months' data; (d) 4 months' data; and (e) Displacement of the highest density locations.

3.5 Conclusion and Discussion

This study examined the effectiveness of LBSM on calculating intra-individual activity space indicators. More specifically, the results revealed how different indicators affect the magnitude of seven intra-individual activity space indicators in three Chinese cities.

First, we uncovered the inherent regularities of individuals' frequently visited locations, network structure among those check-in locations, and possible home locations based on LBSM data. Although individual activity spaces are mainly determined by their frequently visited locations, LBSM users' activity spaces are also affected by the size, spatial distribution, and spatial structures of a city. Moreover, activity space is also influenced by how activity space is conceptualized and measured. There is also a noticeable difference between SDE and convex hull, and it seems that the bigger the city, the larger the size difference is between convex hull and SDE. While comparing the sizes measured by the four external activity space indicators, it is clear that the activity space represented by ROG areas are more stable than the other three indicators in all three study areas. In summary, the ROG method is less sensitive to outliers and covers the most possible activity spaces of LBSM users than SDE. It also proves that the alpha shape indicator describes a more accurate boundary and is less affected by outliers than convex hull.

Second, we discovered the correlation between the change of indicator values and the amount of data used in modeling activity space. The results revealed that all four external and two internal activity space indicators (i.e., entropy and MST) increase when the amount of data increases, eventually, their proportion of increase slows down and approaches zero. However, the home location indicator (i.e., Kernel density) implies that when the data collection duration increases, the displacement of highest density location tends to decrease until it reaches a stable point. Although the displacement decreases when the amount of data increases, the displacement still follows the same pattern as the other indicators which results in stabilization over time. The effectiveness test of LBSM data in modeling human activity space helps to select an appropriate data size in future human activity related studies.

In summary, the results indicated that LBSM users' activity spaces can be presented and measured by different activity space indicators at an individual-level. As the data size increases, the magnitude of defined indicators approaches a steady point. Each of the seven measures represent a methodological variation on evaluating the effectiveness of LBSM data usage in modeling individual activity space.

These intra-individual related indicators together can capture a comprehensive view of individual activity patterns, such as the spatial extent of activity space, the regularity of daily activities, the diversity of movements among POIs, and the structure of movements among their check-in locations. Through the effectiveness test, we can select more effective and reliable activity indicators and data sizes in activity space studies. The choice of appropriate indicators and data sizes to represent individual activity space ensures the integrity, accuracy, and reliability of the activity space modeling. We foresee that the broader impact of this research will yield an enhanced understanding of applying LBSM data in human activity studies and other widely applicable areas of geography, such as transportation and urban planning.

4 TEST THE EFFECTIVENESS OF INTER-INDIVIDUAL ACTIVITY SPACE MODELING

4.1 Introduction

Measuring the similarity between individual activities is an effective way to reveal human dynamics and understand inter-individual variability. The study of activity similarities can identify the relation and commonality of the activity patterns between individuals (Wang et al. 2011, Liu et al. 2014). Traditional human similarity measurements are limited by the difficulty of acquiring enough data sets at an individual-level in a large spatial environment (Yuan and Wang 2018). With the development of location-based technologies, measuring individuals' similarities based on their movements has attracted a lot of attention. Location-based technologies, such as GPS-equipped smart phones, have provided a more flexible way to collect where and when people interact with the environment (Scholz 2018, Carrasco et al. 2008, Liben-Nowell et al. 2005, Eagle et al. 2009). Therefore, these technologies generated valuable datasets for researchers in the field of activity studies (Ahas et al. 2015, Doran et al. 2016, Lee et al. 2016, Resch 2013, Wu et al. 2014, Xu et al. 2015). Among these platforms, LBSM attracts users to share their daily activities with their friends and followers, and provides abundant information of individual-level daily activities (Hemment 2006, Sui and Goodchild 2011, Thielmann 2010, Tuters and Varnelis 2006, Varnelis and Friedberg 2008), and therefore strengthens the role of social media as a proxy to understand human behaviors and complex social dynamics in geographic spaces (Cao et al. 2015). Therefore, LBSM data offers various opportunities for researchers to explore and understand human activity similarities.

However, discovering similar LBSM users' activity patterns is challenging. First, dealing with big datasets like LBSM data is computationally complex. Efficient methods and algorithms are required to undertake the analysis of mining valuable information. Second, it is difficult to accurately model LBSM user behaviors due to data variety and complexity. For example, it is challenging to integrate spatial-temporal data from different LBSM platforms at different levels of completeness. Third, the low-resolution spatial and temporal information from LBSM data brings extra challenges to similarity analysis. In other words, the variability in individual activity patterns makes human activity measurement a challenging task, especially when individual activities differ in both space and time dimensions. Most importantly, unlike sequential GPS data with a high spatial resolution, the sparseness of check-in locations increases the difficulty of discovering similar LBSM user patterns.

Human activity similarity detection relies on analyzing movement patterns. The key for human activity similarity analysis is to find out how to measure the similarities between two activities (Liu and Schneider 2012). Although many studies have analyzed the similarity of user pairs' trajectories (Lv, Chen and Chen 2013, Tiakas et al. 2009, Scholz 2018), there is not sufficient research to analyze and compare the activity similarity of individual activity patterns based on sparse datasets. The main characteristics of point-based LBSM data are the nonsequential and sparse nature of the check-in points, which are very different from the trajectory data. Hence, unlike traditional trajectory similarity studies, this chapter measures LBSM users' activity similarities based on their check-in locations. Chapter 3 explored LBSM activity space by computing and comparing different indicators. However, as discussed by Golledge and Stimson (1997), regularly visited activity locations as well as the travels between and around these locations form an individual activity space. Therefore, it is equally important to conduct activity space studies by comparing the specific point patterns that form the activity space.

To address this problem, there are three steps in this analysis to reach this goal. First, we clustered all LBSM check-ins based on the density-based spatial clustering of applications with noise (DBSCAN) by their geographic coordinates. After clustering, each check-in is assigned with a cluster identifier (i.e., Cluster ID) which shows the spatial group of the check-ins. Second, we summarized the number of check-ins in different clusters for each LBSM user and organize it into a vector space model (VSM) to represent user activity patterns in each cluster. Third, we extended VSM to a spatial-temporal vector space model (ST-VSM) by taking into account their check-in time (i.e., morning, afternoon, and night). We then calculated LBSM users' activity similarities by applying an extended cosine similarity function and evaluated our approach from both spatial and temporal perspectives.

The remainder of this chapter is organized as follows: Chapter 4.2 discusses related work in the areas of inter-individual study and a review of existing activity similarity measurements. Chapter 4.3 introduces the fundamental research design, including the research framework and the main steps of the similarity measurement. Chapter 4.4 presents the experimental results based on LBSM data for three study areas. Chapter 4.5 concludes the research and discusses future directions for studies.

4.2 Measure the Similarity of Activity Patterns

Nowadays, people use LBSM to share their daily life and LBSM platforms provide a larger volume of spatial-temporal data. These location-based data can be acquired through streaming APIs. It brings new possibilities for researchers to analyze human activity patterns from both individual (Musolesi, Hailes and Mascolo 2004) and aggregated perspectives (Bawa-Cavia 2011, Kwan 2000, Mazey 1981). Exploring human activity patterns plays a crucial role in understanding and predicting individual patterns such as activity scheduling, social network structure, and location prediction (Batty 2009, Cullen and Godson 1975). Cho et al. (2011) studied how social constraints such as friendship influence individuals' movements. However, most of these studies haven't considered the spatial-temporal relationship between different LBSM users' patterns.

Both space and time dimensions play an important role in shaping people's access to certain locations. Therefore, space and time are two essential elements that contribute to LBSM user activities (Li et al. 2013). An active area of research in recent years focuses on the influence of space-time constraints on accessibility (Kwan et al. 2003, Ahas et al. 2015, Li et al. 2013, Schönfelder and Axhausen 2016). Individual activity spaces also differ in time because individuals' spatial movements are confined by their daily activity scheduling. It is a complex process that cannot be fully depicted by static spatial information. Many studies of activity similarities focus on static spatial information, without looking into how the understanding of the issues can be enriched through the lenses of time and mobility (Jones and Pebley 2014, Ren, Tong and Kwan 2014, Wong and Shaw 2011). On LBSM platforms, each individual has their unique behaviors in space and time. Therefore, detecting their similarities can be useful for identifying different type of LBSM users' spatial-temporal activities (Kwan 2000). Hence, this study examines the similarity of individual activity spaces from a spatial-temporal perspective.

4.2.1 Similarity Measurements of Trajectory-based Data

As discussed in Section 4.1, activity space studies not only focused on the morphology of the activity space but also the specific point patterns of visited locations—the former measures its basic characteristics and the latter depicts the internal structure of how an activity space forms. This chapter focuses on testing the effectiveness of inter-individual similarity measurements in modeling the activity space by mining similar point patterns that form activity spaces.

One basic type of similarity measurement is the distance-based measurement (Wang et al. 2013), where the distance between trajectories reflects the underlying similarities between the two items. The most commonly used measurement is Euclidean distance (Joh et al. 2002, Joh, Arentze and Timmermans 2001, Liu and Schneider 2012, Buliung and Kanaroglou 2006). It measures the average distance between the corresponding points of two trajectories of the same length. However, the main problem of Euclidean distance is that it is very sensitive to outliers and can not compare the trajectories of different lengths. The dynamic time warping (DTW) method is a robust distance measurement to match stretched or distorted time series (Myers, Rabiner and Rosenberg 1980). It has been used to measure the similarities between individual activity curves in the clustering analysis (Senin 2008, Keogh and Ratanamahatana 2005, Kim, Park and Chu 2004). Levenshtein distance, referred to as the edit distance, is one of the most popular string matching methods for measuring the difference between two sequences (Levenshtein 1966). The edit distance method is used to calculate how dissimilar two trajectories are by counting the minimum number of steps required to transform one trajectory into the other (Chen and Ng 2004, Chen, Özsu and Oria 2005, Scholz 2018). The biggest advantage of the edit distance is that it does not require the two sequences to have the same length. Celik and Dokuz (2018) proposed a similarity measurement based on Levenshtein distance which calculates users' similarities by taking into account both location similarity and the order of locations visited by LBSM users. Yuan and Raubal (2014) extended the edit distance method by incorporating the spatial distribution of cell towers, and applied the developed spatiotemporal edit distance to compare trajectories. The Fréchet distance is a similarity measurement which considers the location and the order of points on trajectory curves (Buchin and Purves 2013). Buchin and Purves (2013) employed equal time distance and Fréchet distance to explore the influence of speed on trajectory similarities. They used space-time prisms to model trajectories and calculated the similarities of these prisms based on Fréchet distance.

Although many studies have analyzed trajectory similarities in various ways, there is not sufficient research to measure the similarities of LBSM users' check-in patterns with low sampling resolution.

4.2.2 Similarity Measurements of Point-based (Nonsequential) Data

Different from high-resolution trajectory data, LBSM check-ins are event-driven data with uneven sampling rates (Wan, Zhou and Pei 2017). LBSM check-in data contains semantic information when a user posts in certain locations. However, they are sparse in space and time. We should employ appropriate methods to process such types of data. Clustering analysis is a primary method to categorize discrete and unordered points. It provides insight regarding the spatial distribution of a dataset and generates unique identifiers for each detected cluster. However, most clustering algorithms like partition-based methods require a predefined number of clusters before conducting the analysis.

Moreover, the parameter settings have a significant influence on the clustering results, so it is necessary to determine appropriate parameters for the algorithm. To sum up, an effectiveness analysis is necessary to demonstrate how sensitive the results are to different algorithm parameters.

Most clustering algorithms belong to two categories: partitioning-based methods and hierarchical-based methods. Partitioning-based clustering simply divides the set of data into non-overlapping clusters. These algorithms iterate until finding the satisfying parameters which properly partition the inner cluster from the outer cluster. Each partition contains a subset of the dataset. Partitioning-based algorithms often contain two steps: (1) predefine the number of clusters (record as *k*), and (2) assign the closest point to a cluster. For example, the classic k-means clustering algorithm belongs to the category of the partitioning-based method. The basic idea is to calculate and find a point that represents the gravity center of one cluster. A hierarchical-based clustering method identifies a set of nested clusters organized as a tree. It does not need to define the number of clusters *k* as in most partitioning-based clustering methods There are two types of hierarchical-based clustering algorithms: agglomerative (i.e., bottom-up) and divisive (i.e., top-down). The agglomerative approach starts with each point as an individual cluster and then merges the closest pair of clusters together at each step. The key operation of agglomerative clustering is the computation of proximity between two clusters. The divisive approach starts with one big cluster which splits into several clusters, and then it continues splitting until only singleton clusters with individual points remain. However, how to choose the appropriate termination condition is a widely discussed research question in hierarchical-based clustering algorithms.

Different from partitioning-based or hierarchical-based methods, density-based methods do not need a predetermined number of clusters or termination conditions. Given a set of locations, DBSCAN algorithm, one of the most commonly used density-based clustering algorithms, can group locations that are close and isolate outlier points in low-density regions (Ester et al. 1996). Moreover, density-based clustering methods can find arbitrarily shaped geographic clusters (Ankerst et al. 1999, Ester et al. 1996, Sander et al. 1998). Hence, it is an appropriate solution to measure individuals' activity similarities based on low-resolution LBSM data. Yuan et al. (2013) proposed a DBSCAN-based clustering method to discover similar users by taking into account both the spatial and temporal dimensions. Moreover, Ankerst et al. (1999) extended the DBSCAN algorithm to process multiple distance parameters at the same time. They constructed the density-based clustering process with respect to different densities to show the clustering structure of the entire dataset.

Another commonly used measurement of inter-individual similarity is vector-based methods such as VSM (Manning, Raghavan and Schütze 2008, McDonald 2000). VSM represents a set of points as vectors in a vector space. A VSM matrix is able to detect user's activity similarities by transforming spatial points into vector format (Mitchell and Lapata 2008). Each vector in a VSM matrix represents a user's activity pattern. The similarity between two users can be calculated by comparing the deviation of angles between these vectors. One important advantage of VSM is that it can compute the similarity between two users even though their trajectories are partially different. Fundamentally, check-ins within the same cluster share the same vector direction. However, this measure suffers from a drawback: two users with very similar patterns can have a big difference based on this calculation simply because one vector is much longer than the other. A standard way to mitigate this problem is to compute the cosine similarity of the two vectors.

4.3 Methodology

As mentioned in Chapter 1, in addition to intra-individual indicators, we are also interested in evaluating the effectiveness of measuring inter-individual activity similarities based on different models. We conduct an effectiveness analysis by comparing a spatial and a spatial-temporal method based on the same 12-month check-in dataset (Figure 4.1).

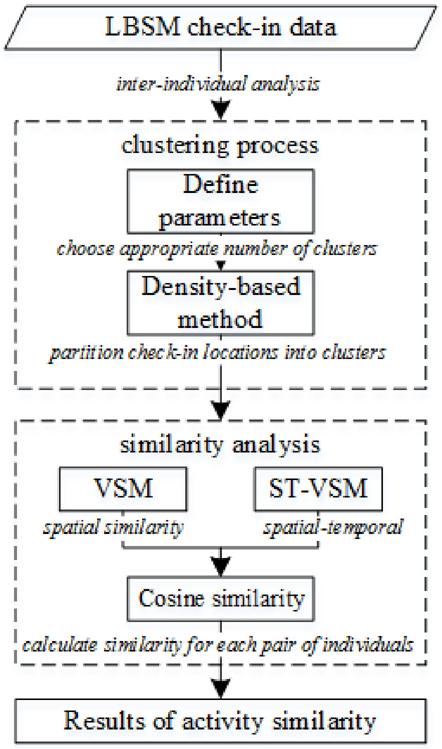

LBSM check-in data
inter-individual analysis
clustering process
Define parameters
choose appropriate number of clusters
Density-based method
partition check-in locations into clusters
similarity analysis
VSM
spatial similarity
ST-VSM
spatial-temporal
Cosine similarity
calculate similarity for each pair of individuals
Results of activity similarity

Figure 4.1 Workflow of Effectiveness Analysis on the Inter-individual Method.

We measured an individual’s spatial-temporal activity similarities from LBSM in the following two steps: First, we clustered user locations based on a density-based clustering algorithm to group similar check-ins. The clustering method used here is the DBSCAN algorithm. This research chose density-based clustering methods for LBSM data, because it has been proven effective on clustering sparse trajectory data with low sampling resolution (Yuan et al. 2013). A lot of partitioning-based or hierarchical-based clustering algorithms are highly sensitive to the number of clusters. Density-based clustering methods like DBSCAN can avoid these issues because they do not need to provide a predetermined number of clusters. Moreover, an advantage of DBSCAN over many other clustering methods is that it can find arbitrarily shaped clusters.

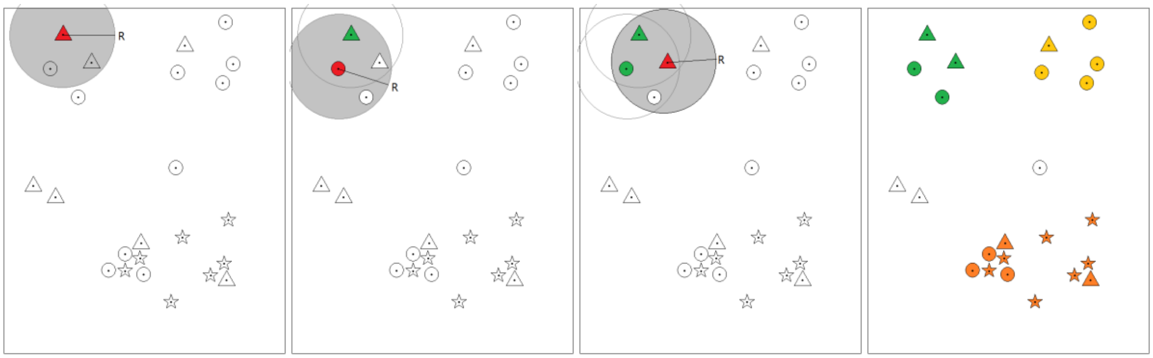
Second, we measured the similarity of individual activities between LBSM users based on the partitioned density surface from spatial and temporal perspectives. After partitioning the points into clusters based on DBSCAN, we measured users' similarity patterns based on the distribution of their check-ins in each cluster. VSM is used to model the spatial clustering distribution of each user. However, ST-VSM considers both spatial and temporal information as two factors to determine the spatial-temporal activity similarity of LBSM users. The experimental results demonstrate the effectiveness of these algorithms in measuring LBSM users' similarity patterns.

4.3.1 Partitioning User Locations into Clusters

(1) The DBSCAN algorithm

The basic idea of DBSCAN is to find clusters based on the density value of spatial locations: the closer the check-ins, the more likely they are in the same cluster. DBSCAN identifies points in high-density neighborhoods as clusters but label points in low-density neighborhoods as outliers.

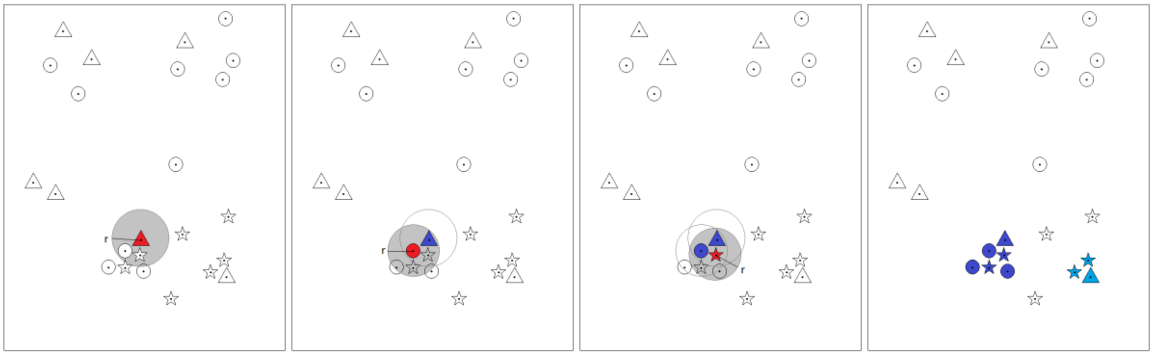
DBSCAN is an efficient spatial clustering algorithm, which can discover clusters of any arbitrary shape and effectively detect noise points. Another advantage of DBSCAN over other spatial clustering techniques is that it does not require a prior number of clusters.



(a) (b) (c) (d)

Figure 4.2 Clustering Analysis Based on DBSCAN (search radius: R; minPts:3)
(a) step1; (b) step 2; (c) step 3; and (d) step 4.

There are two crucial parameters needed in the algorithm: (a) search radius (i.e., *r*) of the neighborhood, and (b) number of minimum points (i.e., *MinPts*) inside of a neighborhood to form a cluster. The DBSCAN clustering algorithm consists of four steps. First, it checks the neighborhood within the search radius of each point. If the neighborhood of the center point *C* contains more than *MinPts* points, a new cluster is created. The center point *C* is called a core point. Second, it iteratively checks whether some of the clusters are density-connected or density-reachable to ensure that all the points that should be in the same cluster are included in that cluster. Any two adjacent clusters sharing one or more points are considered to be density-connected, and any cluster in a sequence of density-connected clusters is defined as density-reachable with respect to any other cluster in this sequence of clusters. Third, the algorithm merges clusters that are density-connected or density-reachable, creating unified clusters with arbitrary shapes. Fourth, the algorithm terminates when no points can be included into a cluster, and there are no connected or reachable clusters. Following this rule, all points within the search radius are added to a cluster and the others are considered as noises (Figure 4.2).



(a) (b) (c) (d)

Figure 4.3 Clustering Analysis based on DBSCAN (search radius: *r*; minPts:3)

(a) step1; (b) step 2; (c) step 3; and (d) step 4.

(2) Reachability distance and clustering structure

As can be seen from the clustering process in Figures 4.2 and 4.3, DBSCAN is highly sensitive to two input parameters that may impact the clustering results.

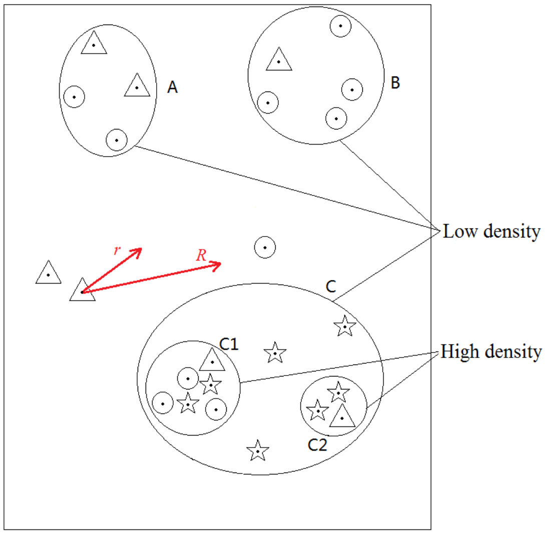


Figure 4.4 Different Densities of Clustering Results by Different Search Radius.

As shown in Figure 4.4, cluster *A*, *B*, and *C* are detected using a set of parameters (*R*,3). However, cluster C_1 and C_2 can be better identified if we apply parameters (*r*,3) to the dataset. If we apply *r* as the search radius, we are unable to detect low-density clusters such as *A* and *B*. Similarly, high-density clusters C_1 and C_2 inside of *C* are overlooked if we apply *R* as the search radius. The main drawback of the DBSCAN method is that the parameter setting of DBSCAN only contributes to one type of results without considering both low-density areas and high-density areas.

In order to overcome this issue, we generated the reachability distance of clusters to explore the clustering structure of the dataset. It creates an ordering of the reachability distance between data points, which identifies high-density clusters first and then continues identifying clusters with lower density (Ankerst et al. 1999). Thus, this method shows the density-based clustering structure of the data. Reachability distance plots can show the relation between clusters of varying densities and corresponding search radiuses. We introduce two definitions to better understand the structure of clusters based on reachability distance:

(a) Core-distance

Here is the definition of the core distance: if point *c* belongs to the dataset, and the point number inside of the nearest neighborhood of *c* meets the *Minpts* requirement, then the core-distance of *c* is the distance from *c* to the nearest neighborhood (red line in Figure 4.5). The range of the core-distance is [0, *search radius*].

$$Core-distance=\begin{cases}Minpts-distance, \textit{points in neighborhood} >= Minpts\\ UNDEFINED, \textit{points in neighborhood} < Minpts\end{cases} \tag{8}$$

(b) Reachability-distance

If point *p* is density-reachable from another point *c*, then the reachability-distance is the larger one between the distance of points *c and p* and the core-distance. The range of the reachability-distance is [*core-distance*, *search radius*].

$$Reachability-distance=\begin{cases}Max\left(core-distance,\ distance\left(C,P\right)\right), & points\ in\ neighborhood>=Minpts\\ UNDEFINED, & points\ in\ neighborhood<Minpts\end{cases} \tag{9}$$

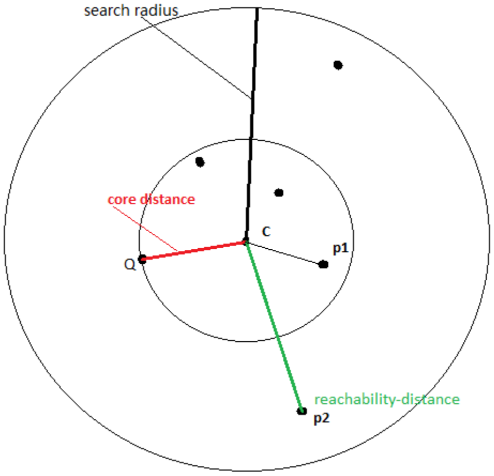


Figure 4.5 Explanation of Core-distance and Reachability-distance.

Based on the definitions of core-distance and reachability-distance, we can retrieve more information about the clustering structure of a dataset.

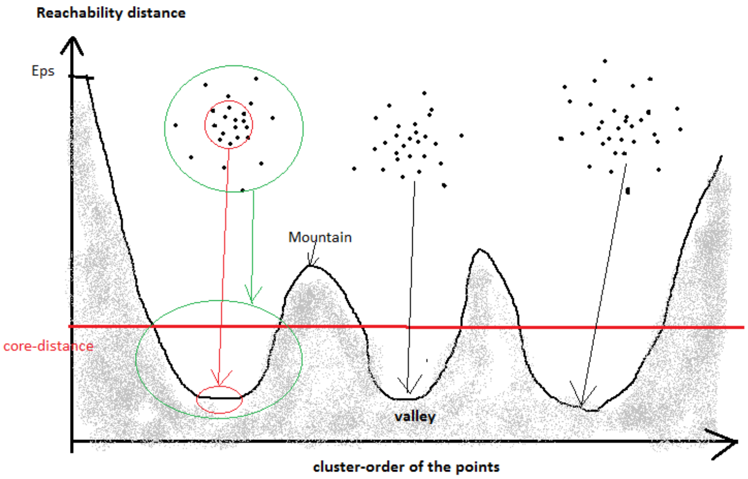


Figure 4.6 A Sample of Reachability-distance.

Figure 4.6 shows several characteristics of the clusters: 1) Points belonging to a higher density cluster have lower reachability distance than their neighbors; 2) Each valley represents one cluster; 3) The deeper the valley, the denser the cluster; and 4) Mountain areas represent the outliers. The higher the mountain, the sparser the points' neighborhood.

(3) Parameter setting

Although density-based clustering methods like DBSCAN do not need to set the number of clusters, it is sensitive to parameters. For example, changing the required minimum number of points in a search neighborhood may unexpectedly affect the clustering results. The reachability distance only provides a generic visualization of the cluster structure; however, we still need to accurately define the size of the search radius. Similar to the measured ratio of between-groups variance against the within-groups variance (Tibshirani, Walther and Hastie 2001), we calculate W_s, which is the average distance between points inside of a cluster, defined as

$$W_s = \sum_{r=1}^{s} \frac{1}{n_r} D_r \tag{10}$$

$$D_r = \sum_{i=1}^{n_r-1} \sum_{j=i}^{n_r} \sqrt{(d_i - d_j)^2} \tag{11}$$

Where s is the search radius, n_r is the number of points in cluster r and D_r is the sum of distances between all points in a cluster. We want to see how W_s changes as the search radius changes.

We use the elbow method to find the appropriate size of the search radius for each study area. We plot the W_s versus the size of the search radiuses to find a visual "elbow" which is the "turning point" of search radiuses. The elbow graph shows what happens to the W_S as the size of search radiuses changes. Moreover, we also calculate the first and second derivative of W_S for each search radius to help choose the appropriate search radius. Specifically, the first derivative represents the changing rate of the indicator W_S and the second derivative represents the rate values of the slope change of W_S. The turning point is when the second derivative reaches a maximum.

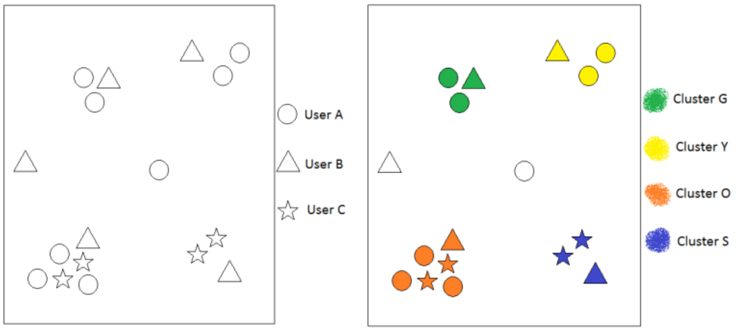
4.3.2 Measure Activity Pattern Similarity based on VSM

Based on the clustering results from the previous step, we labeled each individual user's check-in locations based on the clusters they belong to or as outliers. VSM is originally an algebraic model that measures the relevancy of text documents in information retrieval. In this study, we use it to convert a set of points to vectors in a vector space. It transfers each user's activities into a vector, which is composed of how frequently the check-in points appear in different clusters.

(1) Calculate spatial activity similarity by VSM

The VSM used in this chapter stores the frequency of how many times a user's check-ins appear in each cluster, so each vector represents how a user visits different spatial and temporal clusters.

Figure 4.7 illustrates an example of how VSM works. After applying the clustering method to a dataset, the check-in points of users *a*, *b*, and *c* are assigned to four different clusters. We then calculate the number of check-ins each user has in different clusters and summarize them into a table (Tables 4.1 and 4.2).



(a) (b)

Figure 4.7 (a) Example of User Check-ins; (b) The Clustered User Check-ins.

Table 4.1 The Visiting Frequency of Each User in Different Clusters.

Clusters Users	G	Y	O	S
a	2	2	3	0
b	1	1	1	1
c	0	0	2	2

Table 4.2 The VSM Structure of User *a*, User *b*, and User *c*.

$$VSM = \begin{pmatrix} 2 & 2 & 3 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 2 & 2 \end{pmatrix}$$

(2) Calculate spatial-temporal activity similarity using ST-VSM

As discussed before, it is necessary to combine both space and time when measuring activity similarity between users. However, it is not an easy task to calculate spatial and temporal similarities simultaneously. The main challenge in this activity space study is to compute the spatial-temporal similarities of internal point patterns that form the activity space for LBSM users.

In order to capture temporal similarity patterns, we employed an improved VSM algorithm, ST-VSM, which combines both the spatial and temporal dimensions of user activities. ST-VSM is an extended VSM model to arrange LBSM check-ins into a spatial-temporal vector matrix.

Table 4.3 illustrates an example of how ST-VSM works. We divided user check-ins in each cluster into three groups based on their timestamps (4am-12pm as the morning, 12pm-8pm as afternoon, and 8pm-4am as night). Table 3.5 shows an example structure of ST-VSM. We then investigated the differences of user similarity measurements using the aforementioned methods: density-based clustering methods with VSM and density-based clustering methods with ST-VSM.

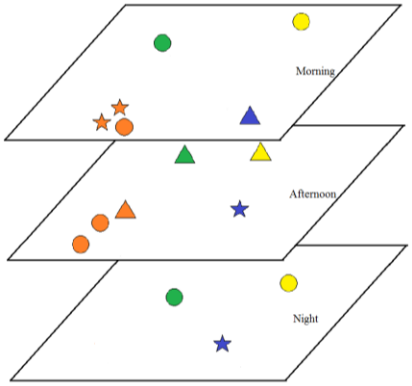


Figure 4.8 The Idea of Create ST-VSM by Timestamp.

Table 4.3 Visiting Frequency of Three Sample Users in Spatial-temporal Clusters.

Clusters	G			Y			O			S		
Users	Morning (M)	Afternoon (A)	Night (N)	M	A	N	M	A	N	M	A	N
a	1	0	1	1	0	1	1	2	0	0	0	0
b	0	1	0	0	1	0	0	1	0	1	0	0
c	0	0	0	0	0	0	2	0	0	0	1	1

Table 4.4 STVSM Records of User *a*, User *b*, and User c.

$$ST-VSM=\begin{pmatrix}\vec{v}_a\\ \vec{v}_b\\ \vec{v}_c\end{pmatrix}=\begin{pmatrix}1&0&1&1&0&1&1&2&0&0&0&0\\0&1&0&0&1&0&0&1&0&1&0&0\\0&0&0&0&0&0&2&0&0&0&1&1\end{pmatrix}$$

The cosine similarity between two vectors is a measure calculating the cosine of the angle between two vectors. This value is ranged between [-1,1]. Here we used the cosine similarity to measure the similarity of activity patterns for each pair of LBSM users in our sample. In other words, we calculated the cosine similarity based on the spatial and temporal distribution of two users, and the similarity value is between [0,1]. The cosine similarity is defined as follows:

$$\cos_{ab\_similarity} = \frac{\overrightarrow{v_a} \cdot \overrightarrow{v_b}}{|\overrightarrow{v_a}| \cdot |\overrightarrow{v_b}|} \tag{12}$$

$$\cos_{ac\_similarity} = \frac{\overrightarrow{v_a} \cdot \overrightarrow{v_c}}{|\overrightarrow{v_a}| \cdot |\overrightarrow{v_c}|} \tag{13}$$

$$\cos_{bc\_similarity} = \frac{\overrightarrow{v_b} \cdot \overrightarrow{v_c}}{|\overrightarrow{v_b}| \cdot |\overrightarrow{v_c}|} \tag{14}$$

Table 4.5 Similarity of Each Pair of Users.

	User *a*	User *b*	User *c*
User *a*	1	0.3333	0.2722
User *b*	0.3333	1	0
User *c*	0.2722	0	1

Table 4.6 Similarity Matrix of Each Pair of LBSM Users.

$$Similarity = \begin{pmatrix} 1 & 0.3333 & 0.2722 \\ 0.3333 & 1 & 0 \\ 0.2722 & 0 & 1 \end{pmatrix}$$

At last, we calculated the cosine similarity for each pair of users and generated a similarity matrix.

4.4 Results

4.4.1 DBSCAN Clustering Results

As mentioned in Chapter 4.3, DBSCAN clustering method requires two parameters: a search radius and the minimum number of points required to form a neighborhood. We use *R* script to implement the reachability distance for three study areas. The argument *minPts* is the minimum number of core points in the search neighborhood, which is often set as the dataset's dimension plus one. Since our point data is two-dimensional, we set our *minPts* to 3 for all three study areas. In this study, because the density of check-ins vary in each study area, we first generated a reachability distance plot for each city to visually inspect the clustering structure (c.f., Figure 4.9).

According to the reachability-distance plots in Figure 4.9, we can see a very different reachability distance distribution within the three cities. Points belonging to a higher density cluster have a lower reachability distance than their neighbors. The density of check-ins in city centers are much higher than other districts in all three study areas. Hence, we separate the study areas into two parts to better capture the clustering structure: the inner municipality area (i.e., the central urban area) and the urban districts outside of the central municipality area but still within the administrative division (i.e., the outer area).

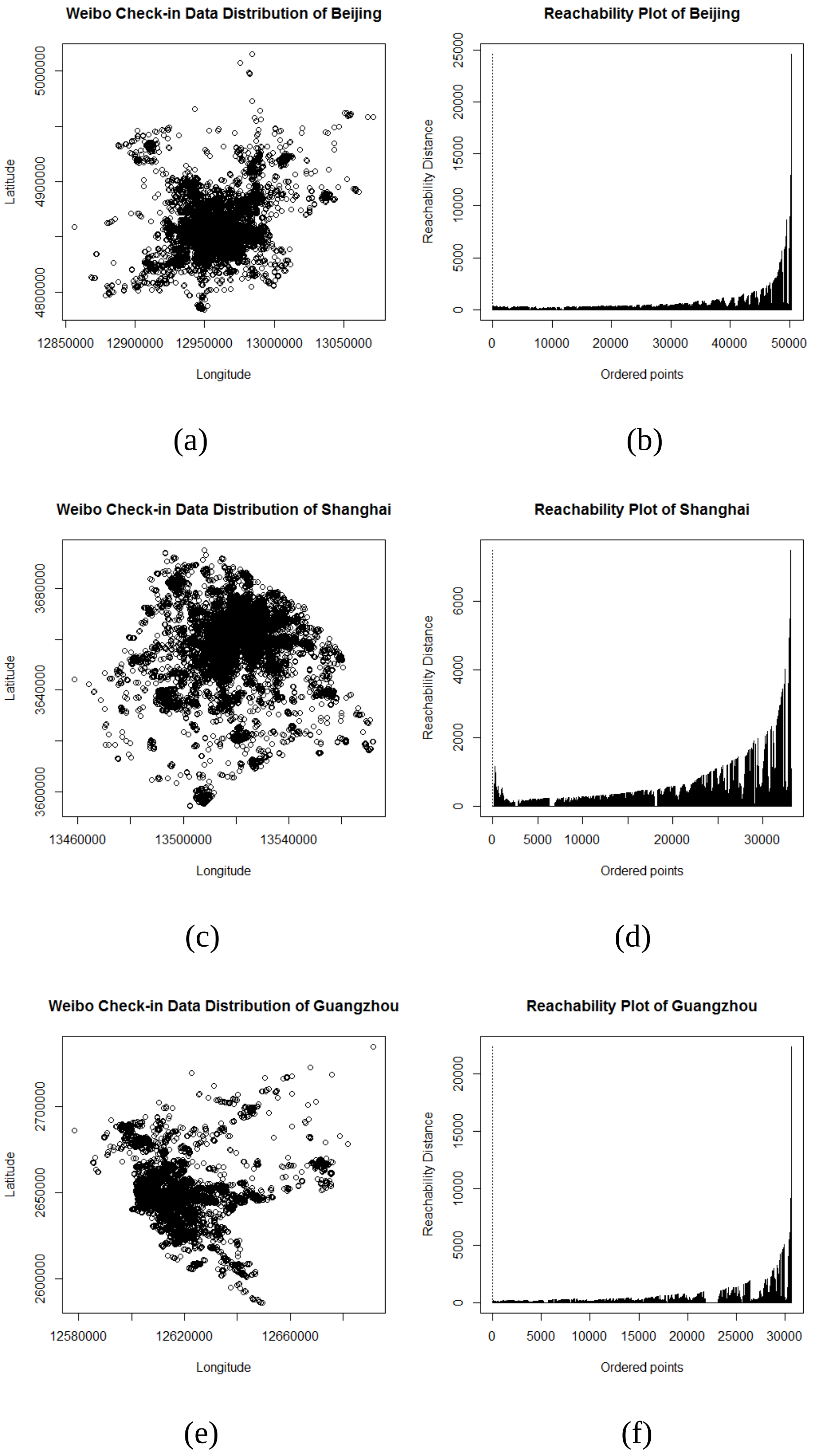
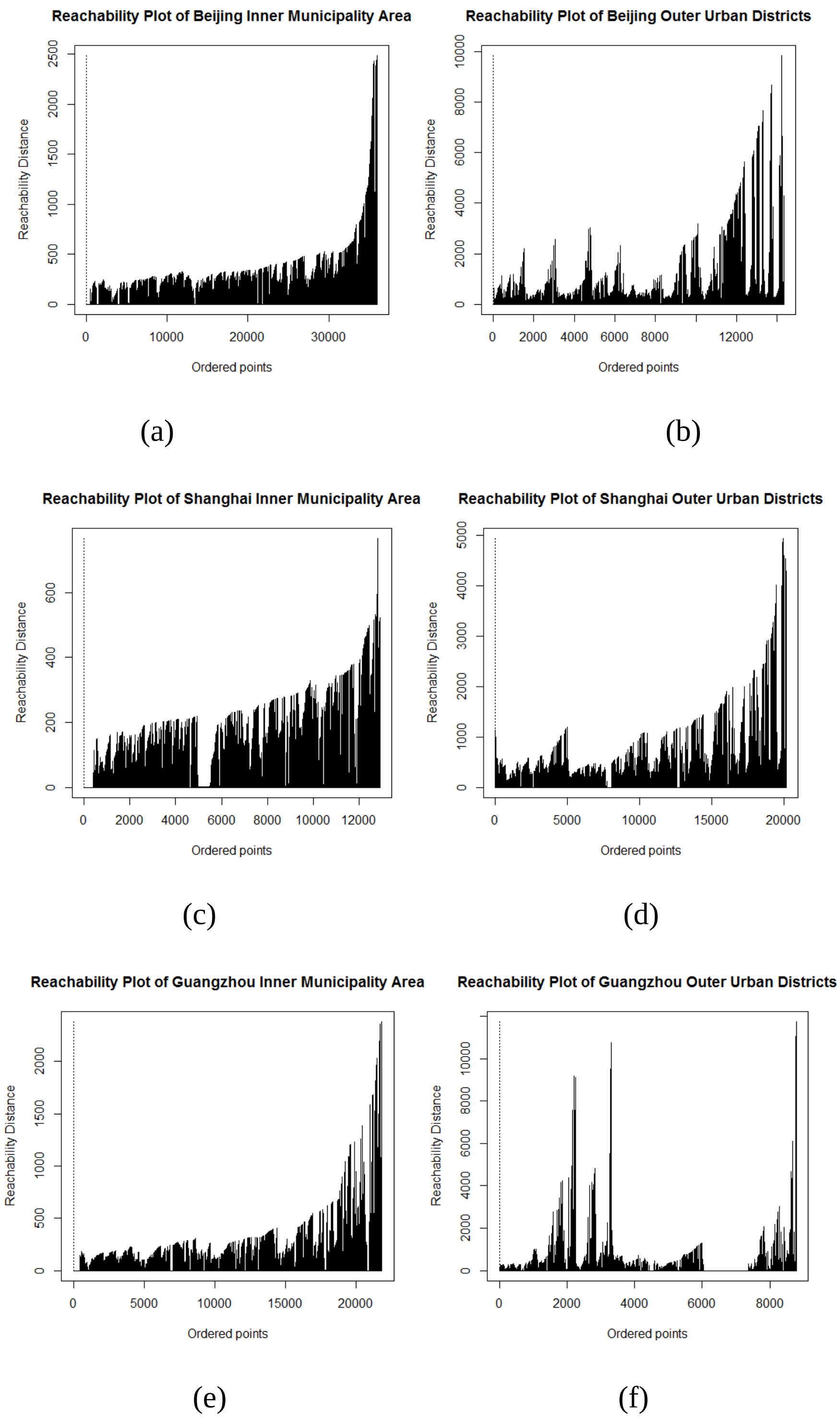


Figure 4.9 Reachability-distance Results:(a) Check-ins in Beijing; (b) Reachability distance of Beijing; (c) Check-ins in Shanghai; (d) Reachability distance of Shanghai; (e) Check-ins in Guangzhou; and (f) Reachability distance of Guangzhou.

Table 2.2 in Chapter 2.3 shows a summary of the total number of check-ins located in the inner municipality areas and outer urban districts in three cities. Beijing and Guangzhou have more than 70% of the check-ins located in their inner areas, while Shanghai only has about 40% of the total check-ins in its inner area. If we apply the same clustering search radius to cluster the entire dataset, it is difficult to detect clustering patterns in the inner and outer areas at the same time. Also, considering the different check-in densities in the inner areas and outer areas, we applied different searching radiuses to them respectively.

Figures 4.10(a), 4.10(c), and 4.10(e) show that most check-ins located in the inner areas are reachable to each other within 1000 meters. And if we choose 500 meters as the search radius, at least 80% of the check-ins belong to the same cluster. Different from Beijing and Guangzhou, the reachability distance of the inner area in Shanghai shrinks to 500 meters. The small size of the inner municipality area in Shanghai is the potential reason why the reachability distance of this region is smaller than the other two cities. The reachability distances between check-ins located in the inner areas are much shorter than the ones in the outer areas (see Figure 4.10). Hence, we chose to use smaller search radiuses for check-ins within the inner municipality areas and larger search radiuses for the outer districts.



(a) (b)

(c) (d)

(e) (f)

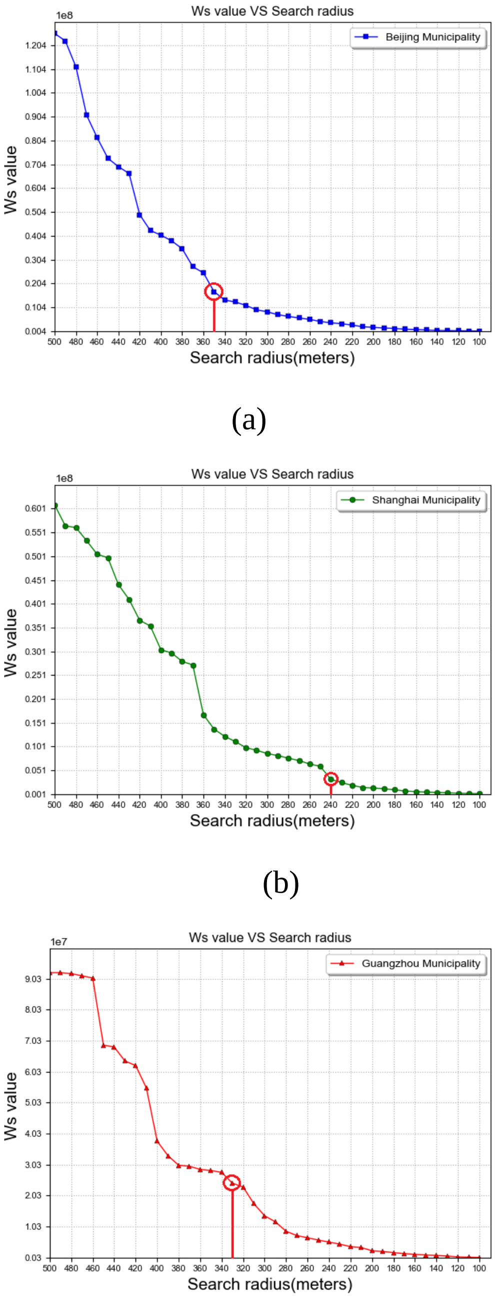
Figure 4.10 Reachability Distance of: (a) The inner area of Beijing, (b) The outer area of Beijing, (c) The inner area of Shanghai, (d) The outer area of Shanghai, (e) The inner area of Guangzhou, and (f) The outer area of Guangzhou.

As a comparison, we plotted the reachability distances of the outer areas for three cities. As can be seen from Figure 4.10(b), 4.10(d), and 4.10(f), a clear clustering pattern appears when we set the search radius around 1,000 meters. There are many "valleys" and "peaks" in the plot. By definition, each valley in a reachability distance plot represents one cluster. The deeper the valley is, the denser the cluster. The outer area of Shanghai shows a different pattern from Beijing. The check-in density decreases from the city center to the outer areas. The further the check-ins are away from each other, the further the reachability distance of the check-ins in the neighborhood are from each other. Guangzhou also shows a different reachability distance plot because of its city morphology. The outer area of Guangzhou is separated into three parts by rivers and hills. As can be seen from Figure 4.10(f), three big clusters are shown around the 400-meter search radius of the outer area in Guangzhou.

According to the previous results of reachability distance in three study areas, we applied different search radiuses for the inner and outer areas respectively using the DBSCAN algorithm. We plotted W_s values with the help of the elbow method (see Figure 4.11) to accurately choose the appropriate search radius for each region.

(1) Parameters Setting of the Inner Municipality Areas

Here is the procedure to deduct the optimal search radius in the inner areas: There are multiple possible turning points in Figure 4.11 and the elbow point (turning point) of W_s occurred between 200 meters and 400 meters for inner areas in all three cities. First, we narrowed down the range manually to select a unique point - because Beijing shows a clear turning point around 350 meters, Shanghai and Guangzhou should have a smaller radius than Beijing because of their city scale.



(a)

(b)

(c)

Figure 4.11 Relations between Search Radius and W_s Values in the Inner Areas:

(a) Beijing, (b) Shanghai, and (c) Guangzhou

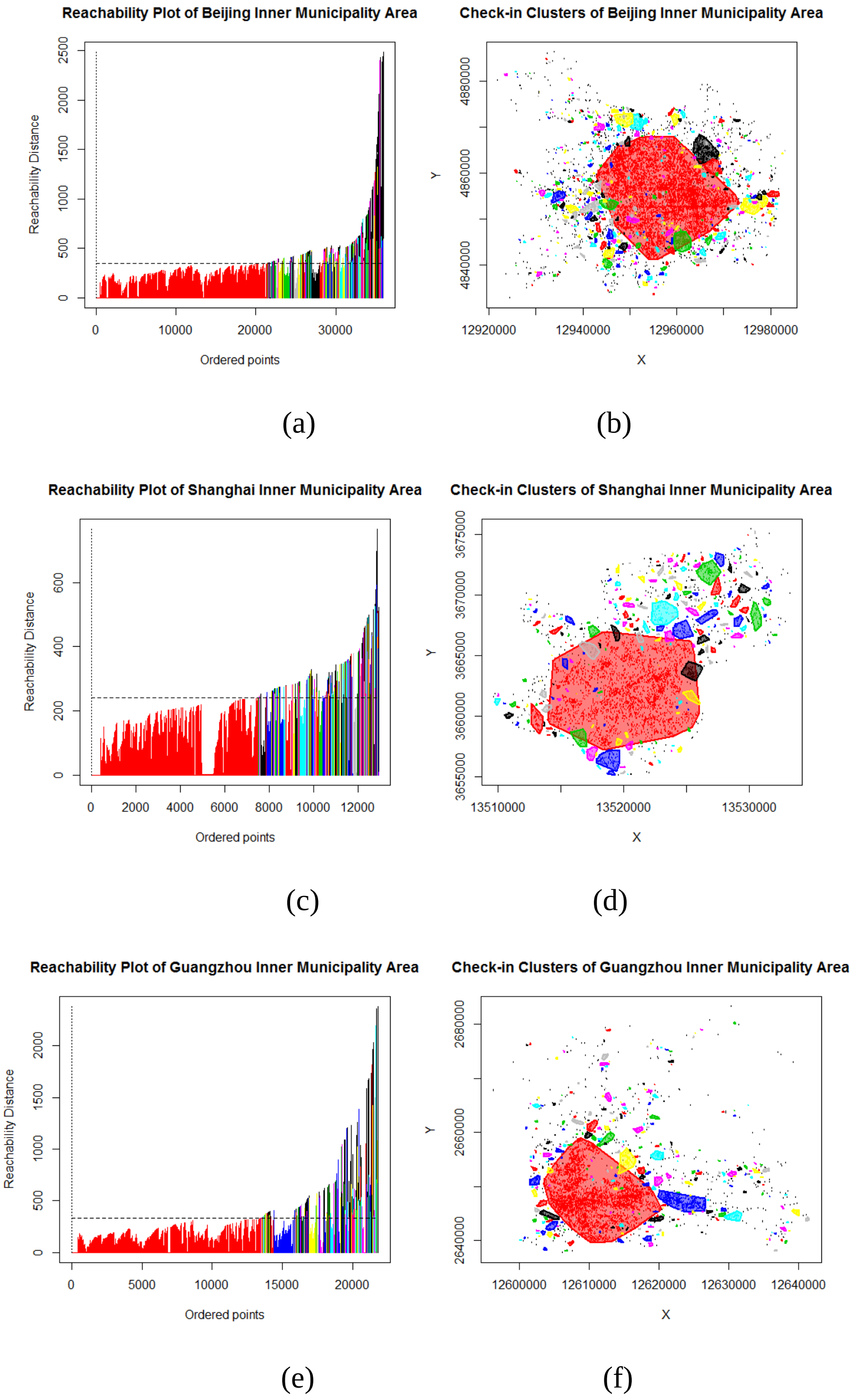
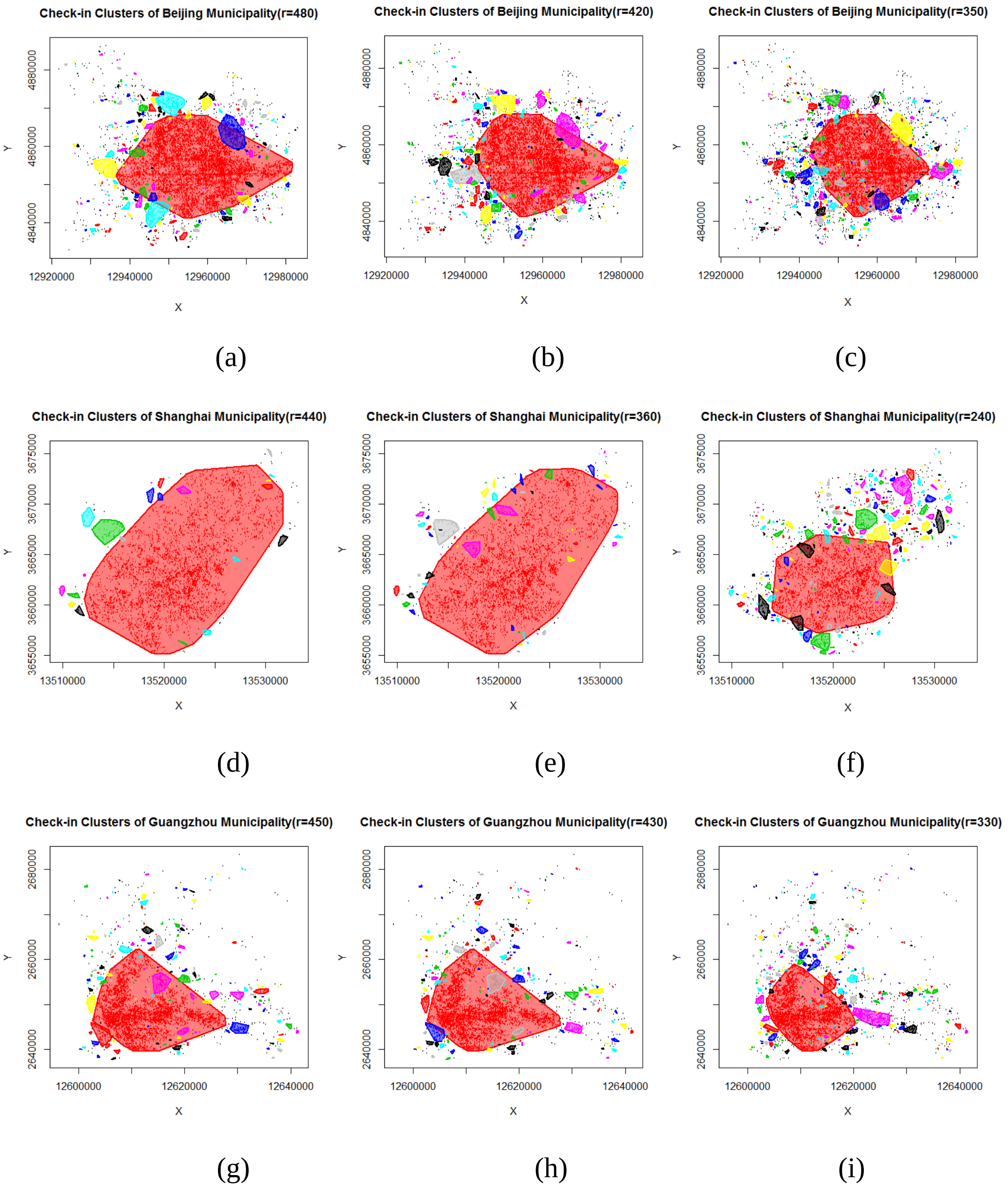


Figure 4.12 Results of the Inner Municipality Area: (a) Reachability distance of Beijing; (b) Clusters of Beijing; (c) Reachability distance of Shanghai; (d) Clusters of Shanghai; (e) Reachability distance of Guangzhou; and (f) Clusters of Guangzhou.

Second, we calculated the first and second derivative of each search radius. The first derivative represented the percentage of increase of the indicator W_s. We calculated the second derivative as the rate of slope change of W_s values. Third, we found the point with the largest second derivative value when r located within the range between 200 meters and 400 meters. In this case, we chose r_{inner_BJ}=350 meters, r_{inner_SH}=240 meters, r_{inner_GZ}=330 meters.

Third, in order to verify the results, we plotted the clusters to inspect them visually (Figures 4.12 and 4.13). To further validate choice of clustering parameters (i.e., search radius) in the inner municipality area for the three study areas, we plotted the clusters based on other potential turning points in Figure 4.11. As can be seen from Figure 4.13, using a larger search radius can potentially lead to over-clustering and underestimate smaller clusters in low-density areas.


Check-in Clusters of Beijing Municipality(r=480)

Check-in Clusters of Beijing Municipality(r=420)

Check-in Clusters of Beijing Municipality(r=350)

Check-in Clusters of Shanghai Municipality(r=440)

Check-in Clusters of Shanghai Municipality(r=360)

Check-in Clusters of Shanghai Municipality(r=240)

Check-in Clusters of Guangzhou Municipality(r=450)

Check-in Clusters of Guangzhou Municipality(r=430)

Check-in Clusters of Guangzhou Municipality(r=330)

(a)

(b)

(c)

(d)

(e)

(f)

(g)

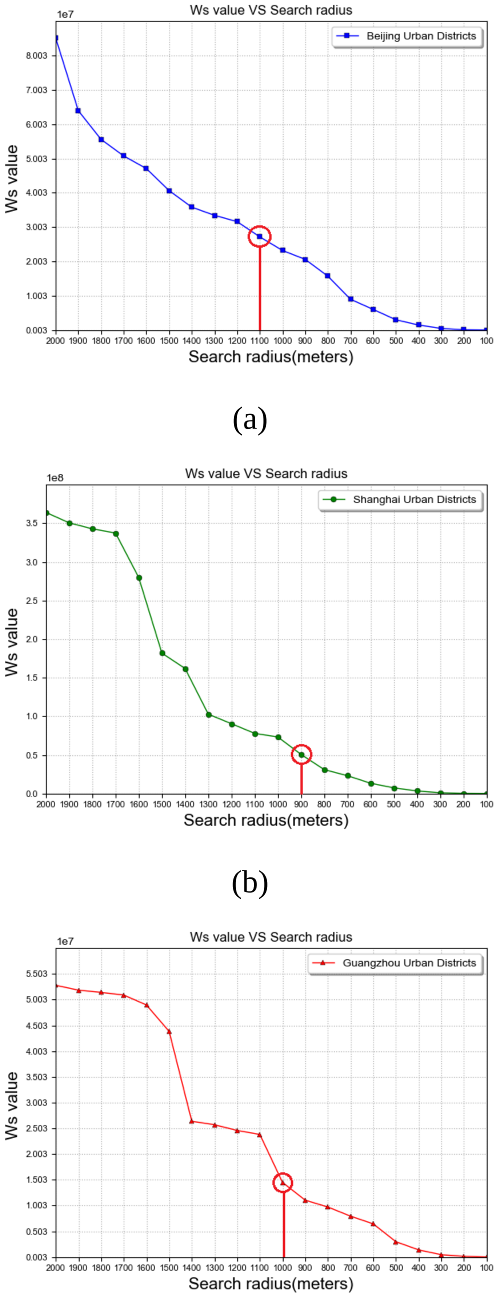
(h)

(i)

Figure 4.13 Clustering Results of the Inner Municipality Area with Different Search Radiuses: (a) r=480m in Beijing; (b) r=420m in Beijing; (c) r=350m in Beijing; (d) r=440m in Shanghai; (e) r=360m in Shanghai; (f) r=240m in Shanghai; (g) r=450m in Guangzhou; (h) r=430m in Guangzhou; and (i) r=330m in Guangzhou.

(2) Parameters Setting of the Outer Urban Districts

We followed the same steps to deduct the optimal radius in the outer areas. We plotted the changing of W_s values in Figure 4.14.



(a)

(b)

(c)

Figure 4.14 Relations between Search Radius and Ws Values in the Outer Areas within: (a) Beijing, (b)Shanghai, and (c) Guangzhou.

Based on Figure 4.14, the elbow point (turning point) of W_s occurs where search radiuses are between 800 meters and 1200 meters for urban districts in all three cities. We then calculated the first and second derivative of W_s and found the point with the largest second derivative value. In this case, we choose R_{outer_BJ}=1100 meters, R_{outer_SH}=900 meters, R_{outer_GZ}=1000 meters. In order to verify the clustering results, we plotted the clusters to verify what they looked like (Figure 4.15-4.16). The clustering results of different search radiuses show that our choice of clustering parameters is appropriate.

In summary, we chose appropriate search radiuses for the inner areas and outer areas for each city respectively by analyzing the relation between W_s and the search radius. According to different check-in densities and search radius values in three study areas, we conducted DBSCAN clustering and the results are shown in Table 4.7.

Table 4 7 Summary of DBSCAN Clustering Results of Study Areas.

Cities	Beijing		Shanghai		Guangzhou	
Summary	the inner	the outer	the inner	the outer	the inner	the outer
number of check-ins	36,014	14,341	12,927	20,195	21,827	8,781
search radius	350	1,100	240	900	330	1,000
number of clusters	211	68	63	86	86	35

We conducted DBSCAN clustering for three study areas based on the optimized search radiuses. The number of check-ins and the clustering results show that the number of clusters in both Beijing and Guangzhou’s inner areas are more than double the amount compared to their outer areas. However, Shanghai shows a different pattern. The number of check-ins and clusters in Shanghai’s inner area is less than the outer area. It implies that the small size of the Shanghai inner area limits the number of POIs and check-ins in the region. After this, we calculated the similarities between LBSM users based on the DBSCAN clustering results.

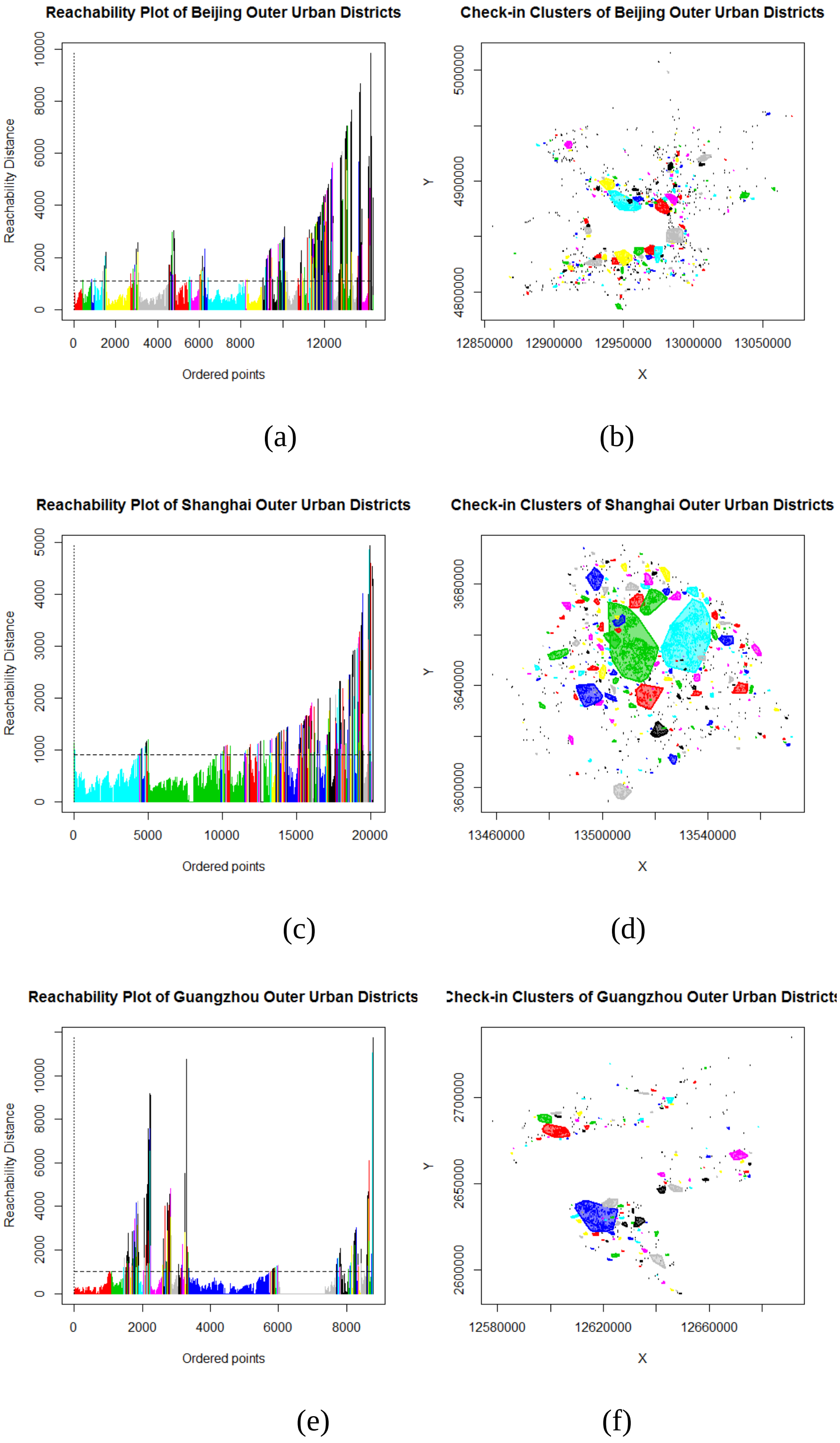
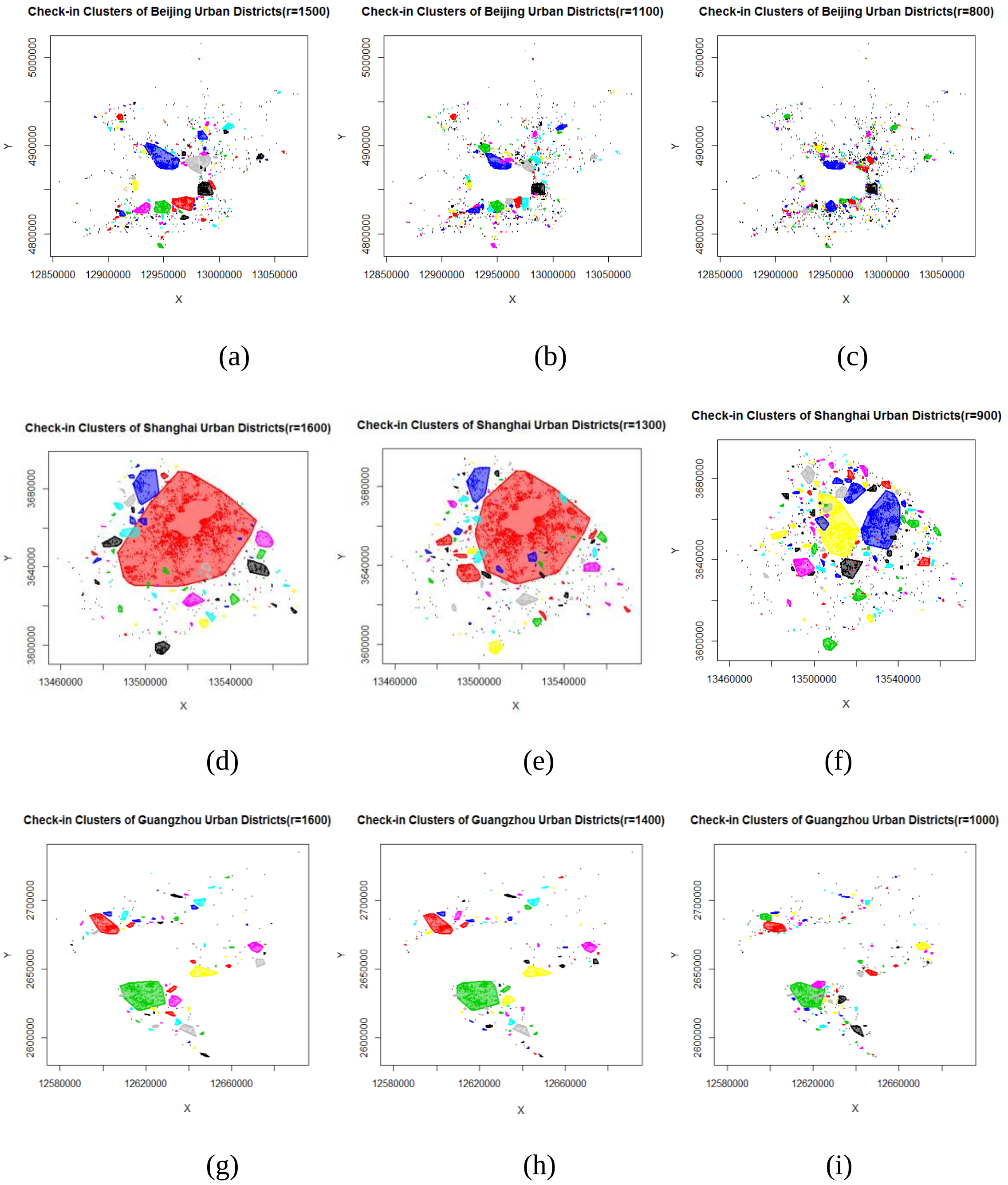


Figure 4.15 Results of the Outer Urban Districts: (a) Reachability distance of Beijing; (b) Clusters of Beijing; (c) Reachability distance of Shanghai; (d) Clusters of Shanghai; (e) Reachability distance of Guangzhou; and (f) Clusters of Guangzhou.



Check-in Clusters of Beijing Urban Districts(r=1500)

Check-in Clusters of Beijing Urban Districts(r=1100)

Check-in Clusters of Beijing Urban Districts(r=800)

(a)

(b)

(c)

Check-in Clusters of Shanghai Urban Districts(r=1600)

Check-in Clusters of Shanghai Urban Districts(r=1300)

Check-in Clusters of Shanghai Urban Districts(r=900)

(d)

(e)

(f)

Check-in Clusters of Guangzhou Urban Districts(r=1600)

Check-in Clusters of Guangzhou Urban Districts(r=1400)

Check-in Clusters of Guangzhou Urban Districts(r=1000)

(g)

(h)

(i)

Figure 4.16 Clustering Results of the Outer Urban Districts with Different Search Radiuses: (a) r=1500m in Beijing; (b) r=1100m in Beijing; (c) r=800m in Beijing; (d) r=1600m in Shanghai; (e) r=1300m in Shanghai; (f) r=900m in Shanghai; (g) r=1600m in Guangzhou; (h) r=1400m in Guangzhou; and (i) r=1000m in Guangzhou.

4.4.2 Activity Similarity Results

Based on the clustering results from the previous step, we labeled each individual's check-in points based on the clusters and time slots they belong to (see Figure 4.17).

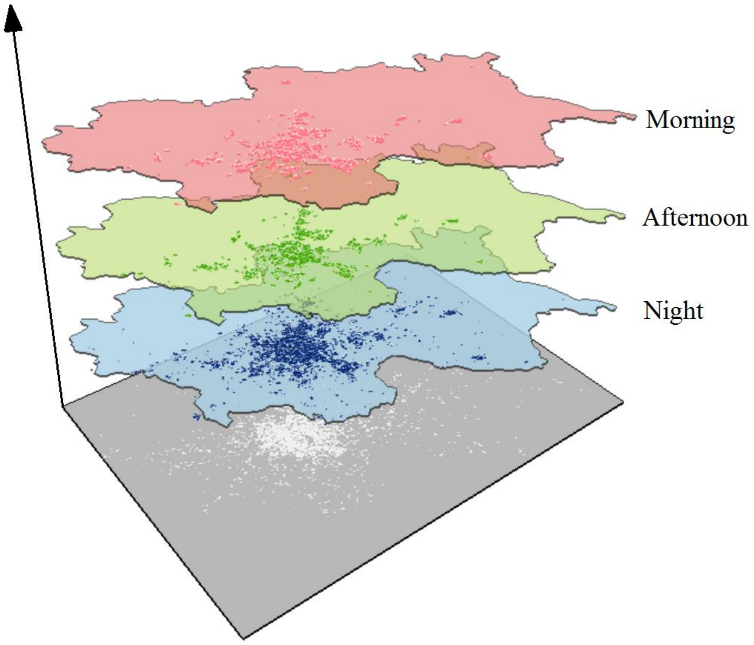


Figure 4.17 Spatial-temporal Structure of Beijing Check-ins.

We used both VSM and ST-VSM to measure the similarities between the individuals' activity patterns as a comparison (Table 4.8). In Beijing, there are over 3.7 million Weibo user pairs that fall into the similarity range [0,0.1], which makes up more than 90% of all user pairs in Beijing. Moreover, it can be noticed that the number of similarity values less than 0.1 occupied more than 84% in all three cities calculated by both VSM and ST-VSM methods. Activity similarity of zero indicates that none of the check-ins from the two users belong to the same cluster. Since ST-VSM further divided check-ins into three different temporal durations, the similarity pattern between two LBSM users was captured at a finer scale. For example, activity patterns happen at the same location but during different times cannot be considered as located in the same cluster. Hence, the number of unsimilar user pairs detected by the ST-VSM clustering method was more than the amount calculated by VSM.

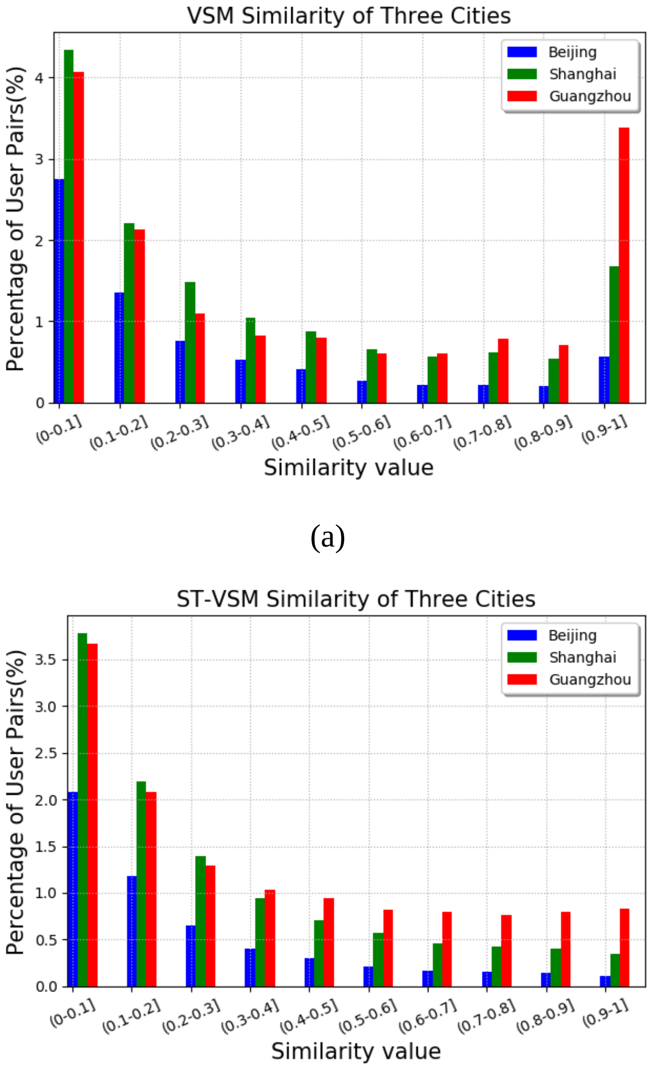
Table 4.8 A Summary of Activity Similarity: (a) Weibo user pairs and (b) Percentage.

Similarity value	Weibo User Pairs					
	Beijing (entire city)		Shanghai (entire city)		Guangzhou (entire city)	
	VSM	ST-VSM	VSM	ST-VSM	VSM	ST-VSM
0	3,633,534	3,707,703	1,355,529	1,399,429	1,140,759	1,167,798
(0,0.1)	107,939	81,397	68,427	59,567	54,696	49,228
(0.1,0.2)	53,375	46,210	34,812	34,524	28,623	27,938
(0.2,0.3)	29,966	25,390	23,264	21,864	14,627	17,280
(0.3,0.4)	20,846	15,812	16,581	14,849	11,119	13,775
(0.4,0.5)	16,168	11,723	13,809	11,126	10,710	12,578
(0.5,0.6)	10,694	8,086	10,340	9,010	8,152	10,965
(0.6,0.7)	8,380	6,766	8,833	7,294	8,065	10,789
(0.7,0.8)	8,117	5,694	9,726	6,698	10,499	10,184
(0.8,0.9)	7,741	5,450	8,465	6,318	9,575	10,713
(0.9,1)	21,840	4,369	26,414	5,521	45,516	11,093

(a)

Similarity value	Percentage Over Total User Pairs					
	Beijing (entire city)		Shanghai (entire city)		Guangzhou (entire city)	
	VSM	ST-VSM	VSM	ST-VSM	VSM	ST-VSM
0	92.7253%	94.6181%	85.9998%	88.7850%	84.9828%	86.9971%
(0-1)	7.2747%	5.3819%	14.0002%	11.2150%	15.0172%	13.0029%

(b)



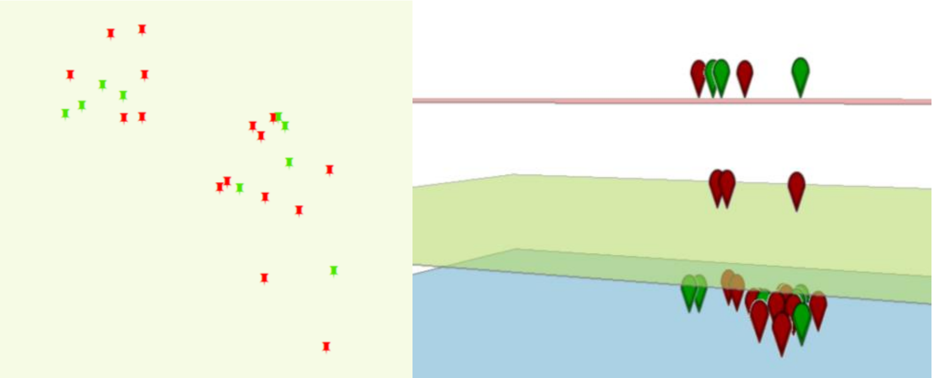
(a)

(b)

Figure 4.18 Similarity Comparison Among Study Areas by: (a) VSM; (b) ST-VSM.

After excluding the zero values in Figure 4.18, the results can be interpreted from multiple perspectives. On the one hand, both the VSM and ST-VSM models used in the study show a power law pattern of similarity distribution in general. The majority of users' activity similarities are lower than 0.1, and only a few users have a higher similarity. On the other hand, the number of similarity values larger than 0.9 calculated by VSM are much higher than the ones from ST-VSM, indicating that although some individuals show a very similar spatial pattern, they have little temporal similarities in their activities. The same analysis can be extended to different cities and different social media platforms to study the impact of spatial and temporal factors on users' activity similarities.

Figure 4.19 shows the spatial and spatial-temporal check-in distribution of two examples users A and B. After applying DBSCAN clustering to their check-in points, we displayed the results from VSM and ST-VSM in Tables 4.9 and 4.10.



(a) (b)

Figure 4.19 Check-in Distribution of Two Users (a) Spatial distribution of check-ins; (b) Spatial-temporal check-ins.

Table 4.9 VSM Structure of User A and User B.

Cluster ID User ID	**2**	**23**	**43**	4	16	22	…
A (127016****)	**4**	**4**	**1**	0	0	0	…
B (263933****)	**6**	**6**	**4**	1	1	2	…

Table 4.10 ST-VSM Structure of User A and User B.

Cluster ID User ID	**2**			**23**			**43**			4			16			22			…
A (127016****)	**1**	**0**	**3**	**2**	**0**	**2**	**0**	**0**	**1**	0	0	0	0	0	0	0	0	0	…
B (263933****)	**0**	**1**	**5**	**2**	**2**	**2**	**0**	**0**	**4**	1	0	0	0	0	1	1	0	1	…

$$\cos_{AB\_similarity_{VSM}} = \frac{\overrightarrow{|v}_{A_{VSM}} \cdot \overrightarrow{|v}_{B_{VSM}}}{|\overrightarrow{v}_{A_{VSM}}| \cdot |\overrightarrow{v}_{B_{VSM}}|} = 0.8389 \tag{15}$$

$$\cos_{AB\_similarity\__{ST-VSM}} = \frac{\overrightarrow{v}_{A_{ST-VSM}} \cdot \overrightarrow{v}_{B_{ST-VSM}}}{|\overrightarrow{v}_{A_{ST-VSM}}| \cdot |\overrightarrow{v}_{B_{ST-VSM}}|} = 0.7070 \tag{16}$$

As can be seen from Table 4.9 and Table 4.10, 43 check-ins of user A and user B are located in three common clusters - clusters 2, 23, and 43. Other check-ins were either located in non-shared clusters or treated as outliers. According to the clustering results, we calculated the spatial and spatial-temporal similarity for this pair of users. The spatial-temporal similarity of user A and B is 0.707, which is lower than their spatial similarity 0.8389. The result demonstrates that though these two users show a very similar pattern in space, they are less similar when we consider the timestamps of their check-in points as they visited the same locations at different times. Therefore, it is necessary to combine both space and time dimensions when measuring individuals' activity similarities.

4.5 Conclusion and Discussion

In this study, we conducted an effectiveness test to measure inter-individual similarities of LBSM users by taking into account both the spatial and the temporal dimensions. We first clustered LBSM users' check-ins by their spatial distribution using DBSCAN clustering method and VSM to describe the inter-individual variables.

The main contribution of this study is to test the feasibility of measuring spatial-temporal similarities of user activities based on sparse LBSM data, as space and time are important aspects to activity space studies. The similarity measurement between individuals allows researchers to discover regular and outlier patterns from LBSM data. The findings indicate that VSM is capable of measuring users' activity similarities in space while the ST-VSM model is effective in describing the spatial-temporal similarities considering both the spatial and temporal dimensions. Note that it is not always necessary to include a temporal component when measuring users' similarities, and researchers can balance the weight of spatial and temporal components based on their practical needs.

The method developed in this study offers an effective approach for assessing activity pattern similarities considering the complexity and multidimensional characteristics of human activities and provides a strategy to identify individuals with similar activity patterns in both the spatial and temporal dimensions.

5 CONCLUSIONS

This chapter provides a summary of the major findings for this dissertation. It also discusses the limitations of this study and topics for future research.

5.1 Summary of Findings and Contributions

Using LBSM data to model human activity spaces has become increasingly popular in the field of human mobility analysis. This dissertation provided a new perspective of applying LBSM check-in data into individual activity space modeling. First, we tested the effectiveness of models in measuring individual activity spaces. Chapter 3 measured intra-individual activity spaces by calculating different external and internal activity space indicators. Second, we tested the effectiveness of models in discovering similar inter-individual activity patterns in Chapter 4. We detected the spatial and spatial-temporal activity patterns by comparing the VSM and ST-VSM methods.

Specifically, Chapter 3 conducted an analysis to evaluate the effectiveness of LBSM data for measuring intra-individual activity space indicators. We obtained internal and external activity space indicators based on different LBSM data sizes from 1 month to 12 months. We demonstrated the differences of using various intra-individual indicators to represent the activity spaces of LBSM users. The results of Beijing, Guangzhou, and Shanghai revealed how these indicators are related to the morphology of the cities. The findings from Chapter 3 indicate that different activity space indicators show different levels of effectiveness in approximating activity spaces based on low-resolution LBSM check-in data. The calculated ROG values were the closest to the approximated limit activity space sizes. It has more than 96% of the observed value over the approximated limit activity space sizes in all three cities using 12-month LBSM data. Moreover, ROG has proved to be a robust external indicator not sensitive to outlier points, which is a common problem of check-in data from LBSM platforms. The findings in Chapter 3 also demonstrated how data collection duration impacts the magnitude of different activity space indicators. As the data size increases, the magnitude of four external and three internal indicators all approach a steady point in three cities. It provided a useful reference to explore a balance point between LBSM data quantity and the accuracy of the analysis.

In chapter 4, we conducted an analysis to evaluate the effectiveness of different models in measuring the inter-individual similarities between LBSM users based on their unevenly distributed check-ins. We clustered check-ins based on the DBSCAN clustering method and compared users' similarities based on both VSM and ST-VSM. We also revealed LBSM users' spatial-temporal activity similarities by considering both the spatial distribution and the timestamps of their check-ins. The results indicate that the ST-VSM method is effective in discovering the spatial-temporal similarities between LBSM users. This extended similarity measurement provided a more robust method to measure user activity similarities based on low-resolution LBSM data. We tested the proposed methods using Sina Weibo dataset in three Chinese cities.

To sum up, the contributions of this dissertation are listed as follows:

This study provided a multilayered research framework to evaluate the effectiveness of LBSM data for activity space modeling from data source, data sampling strategy, and data analytics/modelling perspectives. In other words, the basic structure of this framework is constructed upon modeling human activity spaces using various activity space indicators and models based on different data collection durations from miscellaneous LBSM platforms. Although different LBSM platforms provide different types of information on user activities, this framework helps to extract the most relevant spatial and temporal data to activity space studies and integrate different data structures through the proposed data pre-processing and filtering procedures.

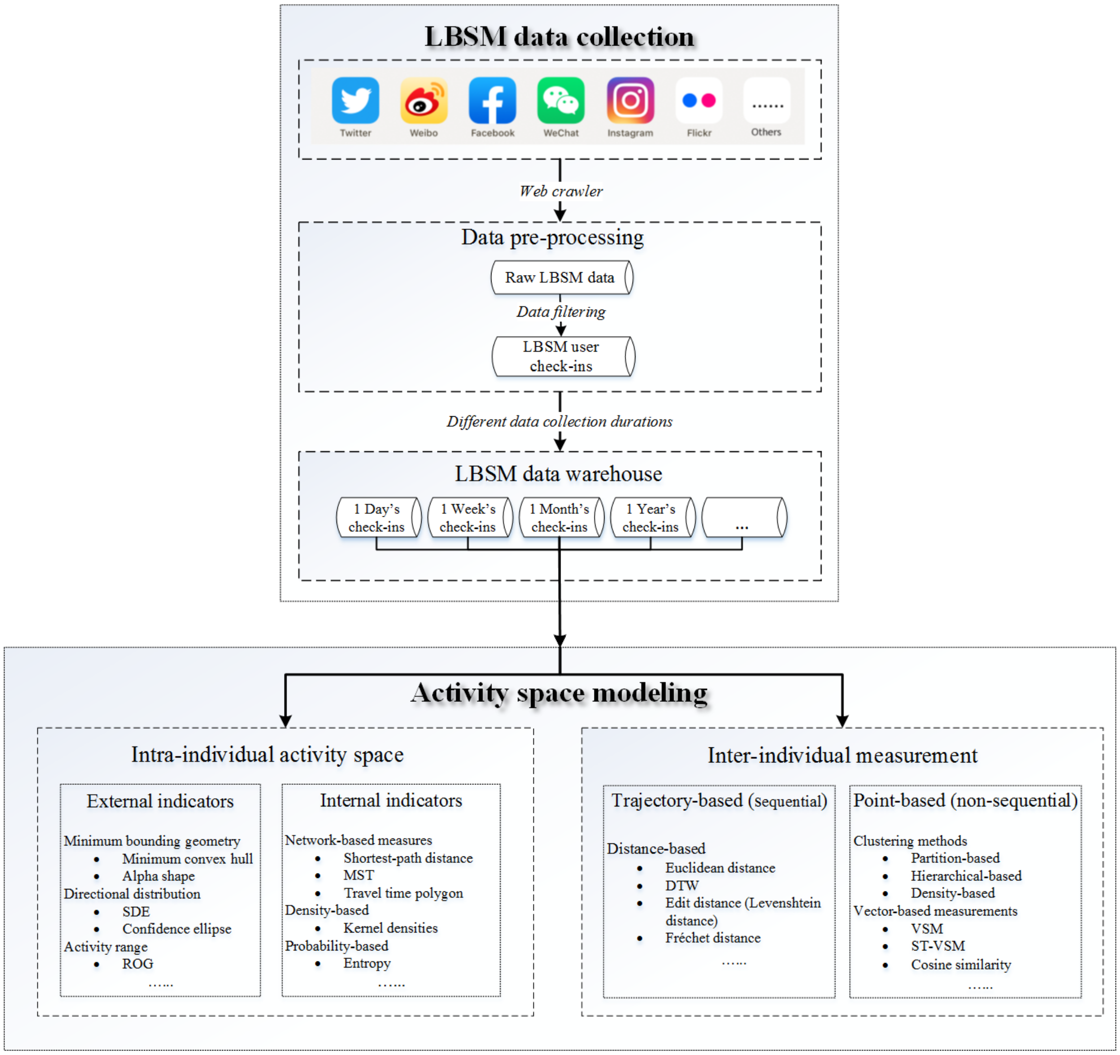


Figure 5.1 Framework for Activity Space Modeling

Furthermore, we also proposed a data sampling strategy which aggregates data collected in different data collection durations. This method can be generalized to discover the correlation among the amount of data used, the stability of activity space indicators, and similarity measurements from different big geo-data, such as mobile phone data and taxi trajectories. Moreover, we applied different indicators to measure the morphology and the internal structure of human activity spaces, as well as the similarity measurement of activity patterns. According to our current framework, researchers can follow the structure and process to conduct their activity space study by customizing the LBSM dataset used, the study areas, the data sampling duration, and activity space indicators and measurements.

In addition to the research framework, this dissertation also contributed to the field from a methodological perspective. Measuring activity space based on LBSM data is computationally intensive due to the complexity and multi-dimensional characteristics of human activities. From a methodological standpoint, this study assessed the effectiveness of low-resolution LBSM data and different activity space indicators in modeling individual activity spaces and measuring activity similarities. The results of this study demonstrate that there is no such thing as "the best indicator" to describe individual activity spaces, because different indicators capture different features of activity spaces (e.g., SDE emphasizes the activity direction and ROG focuses on the distance to the activity center point). Results indicate that the ROG method is more robust and less sensitive to outliers; however, it is not capable of capturing the detailed shape of an activity space like convex hulls or alpha shapes. We demonstrated that these indicators can complement each other to achieve a more complete understanding of individual activity spaces. The ST-VSM method developed in Chapter 4 offers an effective approach for assessing activity pattern similarities considering the complexity and multidimensional characteristics of human activities and provides a strategy to identify individuals with similar activity patterns in both the spatial and temporal dimensions. Our findings proved that testing the effectiveness of LBSM data in modeling activity space provides new insights into the methodological design of activity space studies base on big geo-data, especially for sparse LBSM data with a low spatial-temporal resolution.

The third contribution consists of the various empirical results from the case studies in Chapters 3 and 4. In the intra-individual activity space studies, each of the seven indicators represents diverse characteristic on evaluating the effectiveness of LBSM data usage in modeling individual activity space. As the LBSM data size increases, the magnitude of the defined indicators approaches a steady point. This result demonstrates that all activity space indicators eventually stabilize over time when the data collection duration increases (i.e., amount of data increases). In the inter-individual activity space studies, the results of users' similarities distribution show a power law pattern in all three study areas, where the majority of users show dissimilar patterns and only a few users have a high similarity in their movement. Furthermore, the number of similarity values larger than 0.9 calculated by VSM are much higher than the ones from ST-VSM, indicating that although some individuals show a very similar spatial pattern, they have little temporal similarities in their activities.

To sum up, the case study on three Chinese cities provides a useful reference to explore the balance point between data effectiveness and appropriate sample size from LBSM data. The aggregated activity patterns can provide valuable input for urban planners and policymakers to understand the dynamics of urban residents in three densely populated Chinese cities. The results can be used to optimize the data collection process and to choose indicators in future studies. We foresee that the broader impact of this research will yield an enhanced understanding of applying LBSM data in human activity studies and other widely applicable areas of geography, such as transportation and urban planning. LBSM data also provides valuable information for analyzing inter-personal relations in social sciences, such as community detection, friendship analysis, and anomaly users' behavior detection.

5.2 Limitations

There are several limitations to this study that are worth further investigation.

LBSM data quality issues were not the focus of this study; instead, we focus on the impact of data quantity and data collection durations on activity space modeling. In practice, data quality is an inevitable issue that affects the effectiveness of activity pattern analysis. In fact, the experimental data used in this study is limited because Sina Weibo's platform only provides limited sampling check-ins for all their data to third-party developers (Wang 2015). Moreover, due to the demographic biases of social media users, most active LBSM users are young people who are enthusiastic about new technologies, so the data used in this study is not a randomly selected sample of the entire urban population. It is also possible that computer algorithms instead of real users automatically generate certain Weibo posts. In addition, this study extracts geotagged posts directly based on check-in locations, so we did not differentiate between residents and travelers. Since the scope of this study focuses on introducing a methodology strategy instead of explaining the pattern of residents in a particular city, we did not eliminate users who are potentially travelers.

When discussing how different data sizes affect the results of activity space sizes in Chapter 3.4.2, this study did not consider how data collecting starting time may impact the activity space sizes. For example, in China most people go back home to see their families during the Spring Festival around February each year. It means that people may travel more in February while their activity spaces shrink to a normal size after the holiday. We need a longer data collection duration to test the effect of the starting time point, and future studies can extend this analysis when the data is available.

This study only focused on the spatial and temporal perspectives of human activity patterns not on semantic patterns of human activities. This study did not consider the semantic analysis in modeling human activity space. What people post in certain places and during certain times is highly related to the functionalities of different locations. Understanding the semantics of LBSM posts can further enrich the modeling and analysis of activity patterns.

There was a lack of validation with ground truth data or other LBSM platforms in this study. Even though human activity patterns can be predictable, randomness is still an inevitable component of human mobility (Song et al., 2010), which leads to the difficulties and challenges in ground-truthing human activity studies. The methods and analysis proposed in this study can also be applied to other social media platforms to test their robustness.

In our approach, temporal activities are combined with the spatial dimension in the VSM model. We aggregated timestamps into durations (e.g., mornings, afternoons and nights), which reduced the granularity of the temporal information in the analysis. However, it is possible to treat spatial and temporal dimensions as two independent dimensions in future studies and assign weights accordingly. This allows researchers to gain more insights into human spatial-temporal behavior by prioritizing time or space based on their own practical needs. However, it is computationally heavy to test the weight allocations in different scenarios.

5.3 Future Work

Future studies can explore more activity indicators and combine them with social-economic data, such as the movement direction of activity spaces, networks between home locations and workplaces, accessibility of home location to certain POIs (e.g. hospital, subway station, and school), and relate POIs with land use type. etc.

Another future research direction is generating a more systematic analysis to deal with the uncertainty issues of modeling user activity patterns from LBSM. The methods and analysis proposed in this study can be applied to other social media platforms to test their robustness and extensibility. Although it is challenging to validate activity analysis results with ground truth, future research can take one further step to compare the effectiveness of models in analyzing activity space by making use of the datasets from other LBSM platforms.

Future research can also explore the semantic aspect of LBSM user activities, such as thoughts, emotions, and attitudes expressed on social media. Considering the semantics of check-ins beyond the spatial and temporal perspectives is useful for understanding the purpose of individual activities. Future research can apply the methodology to more application domains of LBSM data, such as breaking news diffusion and criminal detection, etc. The methodology can also be used for more data mining applications such as social network analysis. Extracting users who have a high degree of similarity can be useful for friend recommendation, location recommendation, unusual activity patterns detection, and so on.

Moreover, due to the limitation of data sizes in less developed areas in China, we did not investigate the patterns in smaller cities or rural areas. Future studies can also explore the similarity/dissimilarity between cities in various stages of development when the data becomes available.

LITERATURE CITED

Aggarwal, J. K. & M. S. Ryoo (2011) Human activity analysis: A review. ACM Computing Surveys (CSUR), 43, 16.

Ahas, R., A. Aasa, Y. Yuan, M. Raubal, Z. Smoreda, Y. Liu, C. Ziemlicki, M. Tiru & M. Zook (2015) Everyday space–time geographies: using mobile phone-based sensor data to monitor urban activity in Harbin, Paris, and Tallinn. International Journal of Geographical Information Science, 29, 2017-2039.

Akcora, C. G., B. Carminati & E. Ferrari (2013) User similarities on social networks. Social Network Analysis and Mining, 3, 475-495.

Akkiraju, N., H. Edelsbrunner, M. Facello, P. Fu, E. Mucke & C. Varela. 1995. Alpha shapes: definition and software. In Proceedings of the 1st International Computational Geometry Software Workshop, 66.

Andrew, A. M. (1979) Another efficient algorithm for convex hulls in two dimensions. Information Processing Letters, 9, 216-219.

Ankerst, M., M. M. Breunig, H.-P. Kriegel & J. Sander. 1999. OPTICS: ordering points to identify the clustering structure. In ACM Sigmod record, 49-60. ACM.

Axhausen, K. W., A. Zimmermann, S. Schönfelder, G. Rindsfüser & T. Haupt (2002) Observing the rhythms of daily life: A six-week travel diary. Transportation, 29, 95-124.

Barbosa, M. S., L. da Fontoura Costa & E. de Sousa Bernardes (2003) Neuromorphometric characterization with shape functionals. Physical Review E, 67, 061910.

Batty, M. (2009) Cities as Complex Systems: Scaling, Interaction, Networks, Dynamics and Urban Morphologies.

Bawa-Cavia, A. 2011. Sensing the urban: using location-based social network data in urban analysis. In Pervasive PURBA Workshop.

Becker, R., R. Cáceres, K. Hanson, S. Isaacman, J. M. Loh, M. Martonosi, J. Rowland, S. Urbanek, A. Varshavsky & C. Volinsky (2013) Human mobility characterization from cellular network data. Communications of the ACM, 56, 74-82.

Brockmann, D., L. Hufnagel & T. Geisel (2006) The scaling laws of human travel. arXiv preprint cond-mat/0605511.

Buchin, M. & R. S. Purves. 2013. Computing similarity of coarse and irregular trajectories using space-time prisms. In Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 456-459. ACM.

Buliung, R. N. & P. S. Kanaroglou (2006) Urban form and household activity-travel behavior. Growth and Change, 37, 172-199.

Calabrese, F., M. Diao, G. Di Lorenzo, J. Ferreira Jr & C. Ratti (2013) Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. Transportation research part C: emerging technologies, 26, 301-313.

Cao, G., S. Wang, M. Hwang, A. Padmanabhan, Z. Zhang & K. Soltani (2015) A scalable framework for spatiotemporal analysis of location-based social media data. Computers, Environment and Urban Systems, 51, 70-82.

Carrasco, J. A., B. Hogan, B. Wellman & E. J. Miller (2008) Agency in social activity interactions: The role of social networks in time and space. Tijdschrift voor economische en sociale geografie, 99, 562-583.

Celik, M. & A. S. Dokuz (2018) Discovering socially similar users in social media datasets based on their socially important locations. Information Processing & Management, 54, 1154-1168.

Chapin, F. S. 1974. Human activity patterns in the city: Things people do in time and in space. Wiley-Interscience.

Chen, L. & R. Ng. 2004. On the marriage of lp-norms and edit distance. In Proceedings of the Thirtieth international conference on Very large data bases-Volume 30, 792-803. VLDB Endowment.

Chen, L., M. T. Özsu & V. Oria. 2005. Robust and fast similarity search for moving object trajectories. In Proceedings of the 2005 ACM SIGMOD international conference on Management of data, 491-502. ACM.

Cheng, Z., J. Caverlee, K. Lee & D. Z. Sui (2011) Exploring millions of footprints in location sharing services. ICWSM, 2011, 81-88.

Cho, E., S. A. Myers & J. Leskovec. 2011. Friendship and mobility: user movement in location-based social networks. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, 1082-1090. ACM.

CIW (2018) Weibo monthly active users (MAU) grew to 431 million in Q2 2018.

Cranshaw, J., R. Schwartz, J. Hong & N. Sadeh (2012) The livehoods project: Utilizing social media to understand the dynamics of a city.

Cullen, I. & V. Godson (1975) Urban networks: the structure of activity patterns. Progress in planning, 4, 1-96.

Dijst, M. (1999) Two-earner families and their action spaces: A case study of two Dutch communities. GeoJournal, 48, 195.

Doran, D., K. Severin, S. Gokhale & A. Dagnino (2016) Social media enabled human sensing for smart cities. AI Communications, 29, 57-75.

Duckham, M., L. Kulik, M. Worboys & A. Galton (2008) Efficient generation of simple polygons for characterizing the shape of a set of points in the plane. Pattern recognition, 41, 3224-3236.

Eagle, N., A. S. Pentland & D. Lazer (2009) Inferring friendship network structure by using mobile phone data. Proceedings of the national academy of sciences, 106, 15274-15278.

Edelsbrunner, H., D. Kirkpatrick & R. Seidel (1983) On the shape of a set of points in the plane. IEEE Transactions on information theory, 29, 551-559.

Ester, M., H.-P. Kriegel, J. Sander & X. Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd, 226-231.

Faber, B. (2014) Trade integration, market size, and industrialization: evidence from China's National Trunk Highway System. Review of Economic Studies, 81, 1046-1070.

Fan, Y. & A. J. Khattak (2008) Urban form, individual spatial footprints, and travel: Examination of space-use behavior. Transportation Research Record, 2082, 98-106.

Fayed, M. & H. T. Mouftah (2009) Localised alpha-shape computations for boundary recognition in sensor networks. Ad Hoc Networks, 7, 1259-1269.

GADM (2018) GADM maps and data.

Gesler, W. & D. Albert (2000) Spatial Analysis, GIS, and Remost Sensing Applications in the Health Sciences. Spatial Analysis, GIS, and Remost Sensing Applications in the Health Sciences, 11-38.

Golledge, R. G. 1997. Spatial behavior: A geographic perspective. Guilford Press.

Golledge, R. G. & R. J. Stimson. 1997. Spatial behavior: A geographic perspective. Guilford Press.

Gong, L., X. Liu, L. Wu & Y. Liu (2016) Inferring trip purposes and uncovering travel patterns from taxi trajectory data. Cartography and Geographic Information Science, 43, 103-114.

Gonzalez, M. C., C. A. Hidalgo & A.-L. Barabasi (2008) Understanding individual human mobility patterns. arXiv preprint arXiv:0806.1256.

Gower, J. C. & G. J. Ross (1969) Minimum spanning trees and single linkage cluster analysis. Applied statistics, 54-64.

GSM Association (2003) Permanent Reference Document SE. 23: Location Based Services. Website: http://www. gsmworld. com/documents/lbs/se23. pdf.

Guangzhou International (2016) Population at Year-end

Hasan, S., X. Zhan & S. V. Ukkusuri. 2013. Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In Proceedings of the 2nd ACM SIGKDD international workshop on urban computing, 6. ACM.

Hemment, D. (2006) Locative arts. Leonardo, 39, 348-355.

Huang, L., Q. Li & Y. Yue. 2010. Activity identification from GPS trajectories using spatial temporal POIs' attractiveness. In Proceedings of the 2nd ACM SIGSPATIAL International Workshop on location based social networks, 27-30. ACM.

Järv, O., R. Ahas & F. Witlox (2014) Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone call detail records. Transportation Research Part C: Emerging Technologies, 38, 122-135.

Jiang, B., J. Yin & S. Zhao (2009) Characterizing the human mobility pattern in a large street network. Physical Review E, 80, 021136.

Joh, C.-H., T. Arentze, F. Hofman & H. Timmermans (2002) Activity pattern similarity: a multidimensional sequence alignment method. Transportation Research Part B: Methodological, 36, 385-403.

Joh, C.-H., T. Arentze & H. Timmermans (2001) Pattern recognition in complex activity travel patterns: comparison of Euclidean distance, signal-processing theoretical, and multidimensional sequence alignment methods. Transportation Research Record, 1752, 16-22.

Jones, M. & A. R. Pebley (2014) Redefining neighborhoods using common destinations: Social characteristics of activity spaces and home census tracts compared. Demography, 51, 727-752.

Jones, P. M., F. S. Koppelman & J. P. Orfueil (1990) Activity analysis; State-of-the-art and future directions. New Developments in Dynamic and Activity-Based Approaches to Travel Analysis, 34-55.

Kaisler, S., F. Armour, J. A. Espinosa & W. Money. 2013. Big data: Issues and challenges moving forward. In System sciences (HICSS), 2013 46th Hawaii international conference on, 995-1004. IEEE.

Keogh, E. & C. A. Ratanamahatana (2005) Exact indexing of dynamic time warping. Knowledge and information systems, 7, 358-386.

Kim, S.-W., S. Park & W. W. Chu (2004) Efficient processing of similarity search under time warping in sequence databases: an index-based approach. Information Systems, 29, 405-420.

Kim, S. & G. F. Ulfarsson (2015) Activity space of older and working-age adults in the puget sound region, Washington. Transportation Research Record: Journal of the Transportation Research Board, 2, 37-44.

Kwan, M.-P. (2000) Interactive geovisualization of activity-travel patterns using three-dimensional geographical information systems: a methodological exploration with a large data set. Transportation Research Part C: Emerging Technologies, 8, 185-203.

Kwan, M.-P., A. T. Murray, M. E. O'Kelly & M. Tiefelsdorf (2003) Recent advances in accessibility research: Representation, methodology and applications. Journal of Geographical Systems, 5, 129-138.

Lane, N. D., E. Miluzzo, H. Lu, D. Peebles, T. Choudhury & A. T. Campbell (2010) A survey of mobile phone sensing. IEEE Communications magazine, 48.

Lee, J. H., A. W. Davis, S. Y. Yoon & K. G. Goulias (2016) Activity space estimation with longitudinal observations of social media data. Transportation, 43, 955-977.

Lefever, D. W. (1926) Measuring geographic concentration by means of the standard deviational ellipse. American Journal of Sociology, 32, 88-94.

Levenshtein, V. I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In Soviet physics doklady, 707-710.

Lewin, K. (1951) Field theory in social science.

Li, L., M. F. Goodchild & B. Xu (2013) Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. cartography and geographic information science, 40, 61-77.

Li, Y., Q. H. Li & J. Shan (2017) Discover Patterns and Mobility of Twitter Users-A Study of Four US College Cities. Isprs International Journal of Geo-Information, 6, 17.

Liben-Nowell, D., J. Novak, R. Kumar, P. Raghavan & A. Tomkins (2005) Geographic routing in social networks. Proceedings of the National Academy of Sciences, 102, 11623-11628.

Liu, H. & M. Schneider. 2012. Similarity measurement of moving object trajectories. In Proceedings of the 3rd ACM SIGSPATIAL International Workshop on GeoStreaming, 19-22. ACM.

Liu, Y., X. Liu, S. Gao, L. Gong, C. Kang, Y. Zhi, G. Chi & L. Shi (2015) Social sensing: A new approach to understanding our socioeconomic environments. Annals of the Association of American Geographers, 105, 512-530.

Liu, Y., Z. Sui, C. Kang & Y. Gao (2014) Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data. PloS one, 9, e86026.

Long, J. A. & T. A. Nelson (2013) A review of quantitative methods for movement data. International Journal of Geographical Information Science, 27, 292-318.

Lv, M., L. Chen & G. Chen (2013) Mining user similarity based on routine activities. Information Sciences, 236, 17-32.

Malleson, N. & M. Birkin (2014) New insights into individual activity spaces using crowd-sourced big data.

Manning, C. D., P. Raghavan & H. Schütze (2008) Scoring, term weighting and the vector space model. Introduction to information retrieval, 100, 2-4.

Mazey, M. E. (1981) The effect of a physio-political barrier upon urban activity space.

McDonald, S. (2000) Environmental determinants of lexical processing effort.

Mitchell, J. & M. Lapata (2008) Vector-based models of semantic composition. proceedings of ACL-08: HLT, 236-244.

Mohammady, E. & A. Culotta. 2014. Using county demographics to infer attributes of twitter users. In Proceedings of the joint workshop on social dynamics and personal attributes in social media, 7-16.

Musolesi, M., S. Hailes & C. Mascolo. 2004. An ad hoc mobility model founded on social network theory. In Proceedings of the 7th ACM international symposium on Modeling, analysis and simulation of wireless and mobile systems, 20-24. ACM.

Myers, C., L. Rabiner & A. Rosenberg (1980) Performance tradeoffs in dynamic time warping algorithms for isolated word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing, 28, 623-635.

Na, T., C. Yanwei & K. Mei-Po (2015) Suburbanization, daily lifestyle and space-behavior interaction in Beijing. Acta Geographica Sinica, 70, 1271-1280.

National Bureau of Statistics of China (2017) Population at Year-end by Region.

Phithakkitnukoon, S., T. Horanont, G. Di Lorenzo, R. Shibasaki & C. Ratti. 2010. Activity-aware map: Identifying human daily activity pattern using mobile phone data. In International Workshop on Human Behavior Understanding, 14-25. Springer.

Qu, Y. & J. Zhang. 2013. Regularly visited patches in human mobility. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 395-398. ACM.

Quercia, D., N. Lathia, F. Calabrese, G. Di Lorenzo & J. Crowcroft. 2010. Recommending social events from mobile phone location data. In Data Mining (ICDM), 2010 IEEE 10th International Conference on, 971-976. IEEE.

Ren, F., D. Tong & M. P. Kwan (2014) Space–time measures of demand for service: bridging location modelling and accessibility studies through a time-geographic framework. Geografiska Annaler: Series B, Human Geography, 96, 329-344.

Resch, B. 2013. People as sensors and collective sensing-contextual observations complementing geo-sensor network measurements. In Progress in location-based services, 391-406. Springer.

Sakaki, T., M. Okazaki & Y. Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In Proceedings of the 19th international conference on World wide web, 851-860. ACM.

Sander, J., M. Ester, H.-P. Kriegel & X. Xu (1998) Density-based clustering in spatial databases: The algorithm gdbscan and its applications. Data mining and knowledge discovery, 2, 169-194.

Scholz, R. W. (2018) Space-time modeling of urban population daily travel-activity patterns using GPS trajectory data.

Schönfelder, S. & K. W. Axhausen (2002) Measuring the size and structure of human activity spaces: The longitudinal perspective: ETC 2002-poster session. Arbeitsberichte Verkehrs-und Raumplanung, 135.

--- (2003) Activity spaces: measures of social exclusion? Transport policy, 10, 273-286.

---. 2004. On the variability of human activity spaces. In The Real and Virtual Worlds of Spatial Planning, 237-262. Springer.

---. 2016. Urban rhythms and travel behaviour: spatial and temporal phenomena of daily travel. Routledge.

Senin, P. (2008) Dynamic time warping algorithm review. Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA, 855, 1-23.

Shannon, G. & C. Spurlock (1976) Urban ecological containers, environmental risk cells, and the use of medical services. Economic Geography, 52, 171-180.

Sherman, J. E., J. Spencer, J. S. Preisser, W. M. Gesler & T. A. Arcury (2005) A suite of methods for representing activity space in a healthcare accessibility study. International journal of health geographics, 4, 24.

Silm, S. & R. Ahas (2014) Ethnic differences in activity spaces: A study of out-of-home nonemployment activities with mobile phone data. Annals of the Association of American Geographers, 104, 542-559.

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

Spielman, S. E. (2014) Spatial collective intelligence? Credibility, accuracy, and volunteered geographic information. Cartography and geographic information science, 41, 115-124.

Stefanidis, A., A. Crooks & J. Radzikowski (2013) Harvesting ambient geospatial information from social media feeds. GeoJournal, 78, 319-338.

Steiniger, S., M. Neun, A. Edwardes & B. Lenz (2008) Foundations of LBS. CartouCHe-Cartography for Swiss Higher Education. Obtido em, 20, 2010.

Sui, D. & M. Goodchild (2011) The convergence of GIS and social media: challenges for GIScience. International Journal of Geographical Information Science, 25, 1737-1748.

Susilo, Y. & R. Kitamura (2005) Analysis of day-to-day variability in an individual's action space: exploration of 6-week Mobidrive travel diary data. Transportation Research Record: Journal of the Transportation Research Board, 124-133.

Thielmann, T. (2010) Locative media and mediated localities. Aether: the journal of media geography, 5, 1-17.

Tiakas, E., A. Papadopoulos, A. Nanopoulos, Y. Manolopoulos, D. Stojanovic & S. Djordjevic-Kajan (2009) Searching for similar trajectories in spatial networks. Journal of Systems and Software, 82, 772-788.

Tibshirani, R., G. Walther & T. Hastie (2001) Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63, 411-423.

Tuters, M. & K. Varnelis (2006) Beyond locative media: Giving shape to the internet of things. Leonardo, 39, 357-363.

Varnelis, K. & A. Friedberg (2008) Place: The networking of public space. Networked publics, 15-42.

Veregin, H. (1999) Data quality parameters. Geographical information systems, 1, 177-189.

Wan, Y., C. Zhou & T. Pei (2017) Semantic-geographic trajectory pattern mining based on a new similarity measurement. ISPRS International Journal of Geo-Information, 6, 212.

Wang, D., D. Pedreschi, C. Song, F. Giannotti & A.-L. Barabasi. 2011. Human mobility, social ties, and link prediction. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, 1100-1108. Acm.

Wang, H., H. Su, K. Zheng, S. Sadiq & X. Zhou. 2013. An effectiveness study on trajectory similarity measures. In Proceedings of the Twenty-Fourth Australasian Database Conference-Volume 137, 13-22. Australian Computer Society, Inc.

Wang, R. (2015) Survey on Sina Weibo Research Based on Big Data Mining. International Journal of Data Science and Analysis, 1, 1-7.

Wang, S., J. Min & B. K. Yi. 2008. Location based services for mobiles: Technologies and standards. In IEEE international conference on communication (ICC).

Wong, D. W. & S.-L. Shaw (2011) Measuring segregation: An activity space approach. Journal of geographical systems, 13, 127-145.

Wu, L., Y. Zhi, Z. Sui & Y. Liu (2014) Intra-urban human mobility and activity transition: Evidence from social media check-in data. PloS one, 9, e97010.

Xie, K., K. Deng & X. Zhou. 2009. From trajectories to activities: a spatio-temporal join approach. In Proceedings of the 2009 International Workshop on Location Based Social Networks, 25-32. ACM.

Xu, Y., S.-L. Shaw, Z. Zhao, L. Yin, Z. Fang & Q. Li (2015) Understanding aggregate human mobility patterns using passive mobile phone location data: a home-based approach. Transportation, 42, 625-646.

Xu, Y., S.-L. Shaw, Z. Zhao, L. Yin, F. Lu, J. Chen, Z. Fang & Q. Li (2016) Another tale of two cities: Understanding human activity space using actively tracked cellphone location data. Annals of the American Association of Geographers, 106, 489-502.

Yuan, Y. & M. Medel (2016) Characterizing international travel behavior from geotagged photos: A case study of flickr. PloS one, 11, e0154885.

Yuan, Y. & M. Raubal. 2012. Extracting dynamic urban mobility patterns from mobile phone data. In International Conference on Geographic Information Science, 354-367. Springer.

--- (2014) Measuring similarity of mobile phone user trajectories–a Spatio-temporal Edit Distance method. International Journal of Geographical Information Science, 28, 496-520.

--- (2016) Analyzing the distribution of human activity space from mobile phone usage: an individual and urban-oriented study. International Journal of Geographical Information Science, 30, 1594-1621.

Yuan, Y., M. Raubal & Y. Liu (2012) Correlating mobile phone usage and travel behavior–A case study of Harbin, China. Computers, Environment and Urban Systems, 36, 118-130.

Yuan, Y. & X. Wang (2018) Exploring the effectiveness of location-based social media in modeling user activity space: A case study of Weibo. Transactions in GIS, 22, 930-957.

Yuan, Z., Y. Jiang & G. Gidófalvi. 2013. Geographical and temporal similarity measurement in location-based social networks. In Proceedings of the Second ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems, 30-34. ACM.

Yuill, R. S. (1971) The standard deviational ellipse; an updated tool for spatial description. Geografiska Annaler: Series B, Human Geography, 53, 28-39.