Qingsong Liu, Ph.D., May 2021 GEOGRAPHY

THE ROLE OF MOBILITY IN THE SOCIO-SPATIAL SEGREGATION ASSESSMENT WITH SOCIAL MEDIA DATA (125 pp.)

Dissertation Advisor: Dr. Jay Lee

We are now in an information age. The ubiquitous smartphones and location-aware technologies generate vast amounts of geotagged big data on peoples' spatial activities and movement trajectories. With these geotagged big data, researchers can include population movement into socio-spatial segregation assessment in finer spatial and temporal scales. However, further study is still needed to apply the patterns and structural characteristics of population mobility to research on socio-spatial segregation assessment. Therefore, this research aims to understand and quantify the socio-spatial segregation from activity spaces and human mobility perspective using Volunteered Geographic Information (VGI). I designed a comprehensive analytical framework to evaluate and analyze socio-spatial segregation using human mobility information obtained from VGI. It includes collecting VGI data, classifying VGI users, extracting mobility information from VGI, constructing and analyzing interaction networks, evaluating dynamic socio-spatial segregation, representativeness analysis of VGI, and results' visualization and mapping. This dissertation focuses on analyzing mobility patterns for different VGI user groups, evaluating dynamic socio-spatial segregation, and representativeness analysis of VGI.

Keywords: dynamic socio-spatial segregation, Volunteered Geographic Information (VGI), activity spaces, human mobility.

ANALYZING THE ROLE OF MOBILITY IN THE SOCIO-SPATIAL SEGREGATION ASSESSMENT WITH SOCIAL MEDIA DATA

A dissertation submitted to the

Kent State University in Partial

Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

by

Qingsong Liu

May, 2021

Dissertation written by

Qingsong Liu

B.S., Henan University, 2010

M.S., Henan University, 2013

Ph.D., Kent State University, 2021

Approved by

Dr. Jay Lee, Chair, Doctoral Dissertation Committee

Dr. Xinyue Ye, Member, Doctoral Dissertation Committee

Dr. He Yin

Dr. Ye Zhao

Accepted by

Dr. Scott Sheridan, Chair, Department of Geography

Dr. Mandy Munro-Stasiuk, Interim Dean, College of Arts and Sciences

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENT

The life I have experienced has been a process of cultivating my soul as well as understanding beliefs and trues. Pursuing my Ph.D., I had lost my way. Fortunately, the professors, my family, and my friends have given me their full supports, helps, care, and directions. Without them, I would not have been able to reach here.

I have great pleasure in acknowledging my gratitude to my exceptional committee, comprised of Dr. Jay Lee, Dr. Xinyue Ye, Dr. He Yin, and Dr. Ye Zhao, to provide insightful comments and constructive feedbacks on this dissertation research. I could not have asked for a better committee in my research endeavors. As my doctoral advisor, Dr. Jay Lee and Dr. Xinyue Ye have been especially influential during my study at Kent State University. They inspired me a lot in developing ideas that are fundamental for this dissertation. I want to thank them for being great mentors and collaborators. They are always patient and generous in sharing their valuable knowledge and expertise, which has broadened my perspective on GIScience and space-time analysis.

At the same time, I would like to thank my family members who have always supported me in my studies. My parents have helped me to overcome all kinds of difficulties, both in school and in life. Especially, I would like to give thanks to my wife, Mengmeng Liu. Without her contribution and encouragement, I would not have been able to finish my dissertation.

Finally, I would like to thank my friends, including but not limited to Dr. Adiyana Sharag-Eldin, Dr. Zheye Wang, Weichuan Dong, Zhuo Chen, Patrick Zhao, and many others. Thanks to their presence, my journey has been a colorful one.

CHAPTER 1: INTRODUCTION

1.1. Background

In the past decade, the explosion of mobile devices, location-aware technologies, and social media platforms have accumulated vast amounts of geotagged big data on millions of peoples' spatial activities and movements. These geotagged big data have been used in a variety of applications, such as urban planning (Y. Hu et al. 2015; Gao et al. 2017), public transportation (Iqbal et al. 2014; Järv, Ahas, and Witlox 2014), public health (Hong and Ye 2018), emergency response (Wang and Taylor 2014), and socio-spatial segregation researches (Park and Kwan 2018; Q. Wang et al. 2018; Shelton, Poorthuis, and Zook 2015).

Research on socio-spatial segregation can be dated back to the Chicago school in 1920, which borrowed the ecology terms, such as invasion, succession, and dominance, to describe the dynamics of residential flows and neighborhood composition (Nijman and Wei 2020). With limited ways to collect data, researchers and policymakers relied heavily on census and survey data (Kwan 2009). Thereby, their studies usually focused on socio-spatial segregation in the static residential space (Apparicio et al. 2014). Studies have shown that residential segregation has significant impacts on people's socioeconomic and demographic aspects (Wong and Shaw 2011; Galster and Killen 1995). For example, a high degree of residential segregation is often associated with environmental justice (Crowder and Downey 2010; Jones et al. 2014) and

accessibility to job opportunities (Sultana 2005), educations (Sikkink and Emerson 2008), and medical health services (Williams and Collins 2001; K. White and Borrell 2011).

With the increasing convenience of transportation and communication technologies, there has been a great spatial mismatch between people's residential spaces and their daily activity spaces (Horner and Mefford 2007). People living in the same residential environment can have significantly different moving trajectories and daily life experiences (Park and Kwan 2017a). Different classes of people may be separated in their out-of-home activity spaces (e.g., work, shopping, leisure), which in turn reinforces their disparity in values, choices, lifestyles, and social networks (Wang, Li, and Chai 2012b; Xu et al. 2017). Therefore, segregation studies in residential space that rely on traditional survey data cannot well address the problem of mismatch between residential spaces and out-of-home activity spaces. It is necessary to consider the information of out-of-home activity spaces in socio-spatial segregation studies.

Besides out-of-home activity spaces, the temporal dimension also has important influences on socio-spatial segregation. Conventional segregation studies usually focus on its long-term changes. For example, in the assimilation theory, it usually takes years for minority or disadvantaged groups to move from poorer to more affluent classes (Turner and Wessel 2013). The traditional long-term survey data can meet these research needs. However, many short-term processes also have important spatial effects on socio-spatial segregation (Batty 2002). For instance, people use urban space differently during different times, and their spatial activities are usually repetitive in short terms, such as daily, weekly, and monthly (Silm and Ahas 2014b). Thus, we also need to consider the short-term temporal changes of activity space and population movement in socio-spatial segregation studies. However, traditional survey data are difficult to

help researchers describe these short-term temporal variations. The geotagged big data can capture the timely changes of peoples' activity spaces (Silm and Ahas 2014a; Y. Hu et al. 2015) and their movements between spaces (Yang et al. 2016; Cheng et al. 2020). In addition, the aggregated human movements within one city can reflect interactions between places when citizens become volunteered sensors (Goodchild 2007). The geotagged big data show great potential to fill the above gaps in traditional socio-spatial segregation studies (Yip, Forrest, and Xian 2016; Shelton, Poorthuis, and Zook 2015). Therefore, we can use the geotagged big data to analyze socio-spatial segregation in out-of-home activity spaces and study its dynamic changes.

A large and growing body of literature has explored how to use geotagged big data to understand socio-spatial segregation and disparity in cities. For example, some studies investigate the various activity spaces in segregation (Farber, Páez, and Morency 2012; Wong and Shaw 2011). Other studies focus on the individual-level segregation experience in daily mobility (Park and Kwan 2017b). They have exploited the methodologies to use fine-grained spatio-temporal data to broaden the activity spaces and personal experience in segregation study. But they do not make full use of human mobility information, such as the hierarchical structures of mobility networks and their temporal dynamic changes. This dissertation took a holistic analytical framework to accommodate the aggregated human mobility data in socio-spatial segregation study. This research advanced current segregation research because it revealed the role of mobility networks and their segregation evaluation characteristics.

1.1.1 Activity-Space and Human Mobility In Segregation Assessment

Activity space and human mobility are two interrelated concepts. Activity space describes the spatial extent to which people perform their various daily activities (Horton and

Reynolds 1971). Human mobility refers to the overall pattern of movements that occur between people's different activity spaces. Both activity spaces and human mobility can be analyzed from an individual or a collective perspective (Candia et al. 2008). In this dissertation, I focus on the study from a collective perspective. The collective activity space is the sum of all locations visited by all people in the group. The collective human mobility patterns are the movement patterns of the population group in their collective activity spaces.

Because of data limitations, conventional socio-spatial segregation studies have focused on activity spaces of residence or work. Existing research assumed that segregation is mainly impacted by activities in residential areas or workplaces. For example, their studies only consider people's accessibility to resources distributed in the same area or solely focus on the degree of exposure to other racial/ethnic groups who live in the same area. Conventional socio-spatial segregation studies ignored the out-of-home activity spaces where people engage in routine activities and interact with other groups (Kwan 2009). They especially ignored the critical role of human mobility in out-of-home activity spaces and the temporal dynamic in socio-spatial segregation studies.

First, an analysis of out-of-home activity spaces and mobility patterns is an important part of a comprehensive study of the socio-spatial segregation mechanisms. Many studies have illustrated the need to study segregation in out-of-home activity spaces, closely related to residential segregation and groups' characteristics. There is a disparity in human activity spaces and mobility, constrained by different physical and social conditions, such as ethnicity, class, economic status, education level, etc. For instance, ethnicity plays a significant role in determining some aspects of people's leisure behaviors. African Americans participated more

frequently and incurred higher overall expenditure on the casino trips in a gambling behavior case study (Chhabra 2007). In addition to ethnicity, class and economic status influence the scope and intensity of a person's activity spaces and mobility. Wang et al. (2012) found significant differences in time and space usage between residents inside and outside the so-called privileged enclaves in Beijing. Brands et al. (2014) found that education is most strongly associated with the type of nightlife consumption pursued in two Dutch cities. These studies clearly showed that people with different social statuses spend their time and use urban space differently.

In general, individuals' constraints and perceptions of the city determine the extent of their activity space and thus influence where they are likely to go and who they are likely to interact with (Wang et al. 2018). In turn, the people and objects that individuals contact with also impact or determines their constraints and perceptions of the city (Galster and Killen 1995). This self-reinforcing system implies that segregations of out-of-home activity spaces may exacerbate their socio-spatial segregation in their residential space (Galster and Killen 1995). Therefore, understanding out-of-home activity spaces and mobility patterns is an integral part of understanding socio-spatial segregation among different groups.

Second, socio-spatial segregation also has temporal dynamic changes. The socio-spatial segregation state between different groups can also be affected by temporal changes of activity spaces and mobility patterns. Cities can be considered clusters of 'spatial events' that occur in different times and spaces (Batty 2002). Different areas of a city have very different functions, and they are used at different times of the day (Bromley, Tallon, and Thomas 2003). For example, a bar area is used for leisure activities, and it is often visited by young people during

the nights. Besides, the spatial events are usually repetitive in the temporal dimension (Bromley, Tallon, and Thomas 2003). For example, a nightclub on every Saturday may be full of young people, while having much fewer people on every Monday. The repetitive patterns reflect the spatial and temporal characteristics of people's use of different types of urban spaces and facilities. They are also an important factor in studying socio-spatial segregation. Therefore, to provide a valid measure of socio-spatial segregation, we need to consider temporal dimensions of activity space and group movement in our analysis.

1.1.2 Activity-Space and Human Mobility With Social Media Data

In conventional socio-spatial segregation studies, travel diaries were usually used to construct activity space and human mobility. These travel diaries recorded detailed information about the recorder's activities over the course of a day or several days. They can also provide activity information for a specific group, such as in Kwan's (2008) study of environmental fears among Muslim women, in which 37 respondents were interviewed, and detailed information was collected about each individual's activities and feelings. With these type of data, researchers can explore differences in activity spaces and mobility patterns between different population groups (Kwan 1999), study the spatial characteristics of daily household activities and transportation behavior (Buliung and Kanaroglou 2006), and portray the segregation from the activity space perspective (Wong and Shaw 2011).

However, travel diary surveys usually have a smaller number of subjects, have short survey time, and are costly in money and time. With advances in information and communication technology (ICT) and computing power, an increasing number of new data sources are emerging. The emerging new data makes up for the shortcomings of traditional

survey data to some certain extent. However, we must emphasize that new data is an enrichment, not a replacement, for traditional data. The emerging new data includes Volunteered Geographic Information (VGI) (e.g., Twitter, Foursquare, Sina Weibo) (Baginski, Sui, and Malecki 2014; Yin et al. 2017; Huang and Wong 2015a), phone location data (e.g., call detail record) (Zhao et al. 2016; Toole et al. 2015; Shi et al. 2015), vehicle movement data (e.g., private vehicle trajectory, taxi trajectory) (Tang et al. 2015; Zhu et al. 2017), and public transit system data (e.g., subways and buses) (Chen, Chen, and Barry 2009; Y. Long et al. 2016). As a complement to traditional survey data, these new data have been used to construct activity spaces and describe human mobility.

Among the different new data sources, VGI is mostly used to study activity spaces and human mobility (Liao et al. 2018). In contrast to travel diaries, VGI is generated voluntarily by users. It usually includes a large number of users for a relatively long observation period, and it has a relatively low cost of money and time. VGI provides us people's fine-grained spatial and temporal personal digital footprints in cyber space. It can make up for the shortcomings of traditional survey data in describing users' spatio-temporal activities or trajectories. Although VGI data can provide a comprehensive view of user activity scope, it is not collected to support the analysis of a particular activity space (Liao et al. 2018). In addition, we need to carefully consider its data quality when using VGI to study activity and mobility patterns. Using VGI to infer people's real personal activity spaces and trajectories in the physical world has become a hot research topic. Many studies develop different methods to construct human activities and interactions (Hawelka et al. 2014; Huang and Wong 2015b; Hu, Li, and Ye 2020).

Along with the abundance of individual activity and trajectory data, researchers evaluating segregation issues have begun to shift their socio-spatial segregation studies' focus from residential spaces to out-of-home activity spaces, from a collective level to an individual level, and from a static perspective to a dynamic perspective. These three shifts are not independent but intertwined with each other. For example, when expanding from a static residential space to other activity spaces, researchers have to pay attention to the dynamics of people's activity space and their movements. Thus they have to consider individual-centric and mobility perspectives of socio-spatial segregation. However, most studies ignored or downplayed the role of mobility in the collective sense in segregation evaluation. Given this, researchers have not developed a comprehensive framework for adding mobility to the socio-spatial segregation evaluation yet. This dissertation aims to design a comprehensive framework to evaluate socio-spatial segregation from a dynamic perspective, using mobility data extracted from VGI. This research can advance socio-spatial segregation study, contribute to better urban planning and sustainable development within cities, and provide complementary coordination between cities.

1.2. Research Objectives

This dissertation aims to understand and quantify the socio-spatial segregation from entire activity spaces and human mobility patterns perspective using VGI. I designed a comprehensive analytical framework to evaluate and analyze socio-spatial segregation using human mobility information obtained from VGI. It includes collecting VGI data, classifying VGI users (e.g., visitors and residents), extracting mobility information from VGI, constructing and analyzing interaction networks, evaluating dynamic socio-spatial segregation, representativeness analysis of VGI, and results' visualization and mapping. This dissertation research mainly

focuses on analyzing mobility patterns for different VGI user groups, evaluating dynamic socio-spatial segregation, and representativeness analysis of VGI.

Socio-spatial segregation has many dimensions, the most important of which are spatial evenness (or spatial clustering) and spatial exposure (or spatial isolation) (Reardon and O'Sullivan 2004). First, I studied the spatial evenness of social segregation in Chapter 2. I also investigated how to incorporate movements of residents and visitors into the spatial evenness index. Research in Chapter 2 also analyzed and compared the mobility patterns of different VGI user groups due to their movements' heterogeneity. It provides supports for the following socio-spatial segregation analysis. Second, Chapter 3 studied the methodology to measure the spatial exposure dimension of socio-spatial segregation based on people's mobility patterns. Specifically, I wanted to integrate interaction weights, temporal changes, and hierarchical structure information inferenced from human mobility patterns into the conventional spatial exposure index. Third, I studied and analyzed the representativeness of VGI flow data due to its innate representativeness biases in Chapter 4. These three studies ensure that I can get a better understanding of the generalizability of our research results. Details of each research topic are illustrated below.

First, it was necessary to study people's mobility patterns before analyzing their socio-spatial segregations. Studies have shown that different groups of people have different activity spaces and mobility patterns (Schönfelder and Axhausen 2003; Wang and Zhou 2017). To analyze the mobility patterns of different groups of people, I need to construct their interaction networks based on VGI. A preliminary research question I want to ask is: How to use VGI to construct interaction networks for different population groups from a collective perspective? And

then how to evaluate and measure the interaction networks between different groups and compare their differences? It is difficult to evaluate and measure the interaction networks. Their temporal changes (daily, quarterly, yearly) make the study more complex. Therefore, I need to find a good way to evaluate interaction networks between different groups using VGI data. The first part of my research answers the above questions. It classifies VGI users into different groups and then compares the differences in their activity space and movement patterns.

RQ1: How to uncover different groups of visitors' mobility patterns? What are the differences in mobility patterns between residents and visitors?

I used Twitter data in the study because it was the most typical type of VGI, and it had received the most attention from researchers (Liao et al. 2018). The designed analytical framework included the basic operation to identify whether a Twitter user was a local resident or a visitor. The framework provides support for all the following parts of research in my dissertation. I reviewed all current residential identity detecting methods and selected the most suitable method for Twitter data. The Twitter users were classified into four groups, including local residents, state visitors, US visitors, and international visitors. I then examined activity space differences between the resident group and the other visitor groups from both distribution and interaction perspectives. Based on the first study, researchers can evaluate spatial and temporal differences in interaction networks and use these dynamic interaction networks to evaluate socio-spatial segregation more precisely. Ultimately, these efforts can provide a meaningful reference for policy making.

As mentioned early, due to the mismatch between people's residential space and their out-of-home activity spaces, it was necessary to consider population movement in our

segregation study. But there was no clear way to incorporate this new information. It was important to study methodology to integrate further information into conventional methods/indices. Because in this way, the research results can be seen as a continuation and update of the conventional results. They can also be easily compared with previous research results. Therefore, in the second research topic, I studied integrating new VGI information into the conventional socio-spatial segregation index. I found that the conventional segregation index underutilized the flow network's attributes (e.g., interaction weights, network structure) and the temporal information of flows**.** Thus, this study wanted to integrate interaction weights, temporal changes, and hierarchical structure of flow networks into conventional socio-spatial segregation index. By utilizing this new information, I evaluated socio-spatial segregation more accurately.

RQ2: How can we incorporate flow patterns into conventional segregation measurement with minimal modifications of conventional formulas? Where and when is the new segregation index significantly different from the conventional one? What factors cause their differences?

The second research topic focuses on integrating population flow networks into the spatial exposure index of socio-spatial segregation. Spatial exposure refers to the degree of potential contact, or the possibility of interaction, between two group members in their surrounding environments (Massey and Denton 1988). Much of the literature has used the interaction between regions to improve the conventional spatial exposure index. However, due to data limitations, the interaction strength between regions is only evaluated based on distance, geometry, and boundary information. Regardless of the complex mathematical formula of these methods, their results are still artificial. Therefore, we need to find a more realistic way to assess regions' interactions based on population flow networks and their structural characteristics. The

conventional segregation evaluation index was improved using the temporal dynamics and the flow network's hierarchical structure in my study. In order to analyze the new information's effect on the exposure index, a large number of simulations were also carried out to compare the improved segregation index with the conventional index. Based on the comparison results, how the population flow network impacts the spatial segregation evaluation can be seen. The revealed impacts can be a meaningful reference value for extending other segregation indices.

Both of the above studies focused on exploring how population flow data from VGI can enrich conventional segregation analytics. The default premise in this research is that flow data can represent the movements of the background population. However, there is much debate about whether the new flow data can represent the studied population's movements (Boyd and Crawford 2012). Because it does not have a systematic collection design and is not generated or collected to address a specific question (Liao et al. 2018), some researchers have found high correlations between social media data and other data sets (Lenormand et al. 2014). However, these studies focused on the comparison between two or more mobility datasets. And they did not systematically analyze associations between the representation of mobility extracted by VGI and local socioeconomic factors. Our third research topic aims to fill this gap. This study contributes to the research of representativeness of VGI and socio-spatial segregation. First, representativeness studies are the backbone of all studies that use new mobility data. We can get more solid research results using new data with a comprehensive understanding of its strengths and weaknesses. And the research results can also be more easily generalized and applied broadly into other areas. Second, the representativeness bias of VGI also associates with social inequality. The social-economic inequality of VGI users causes the representativeness issue. This inequality of participation in the social platform (virtual space) reflects the disparity of their

demographic/social-economic conditions in the real world. For example, Li et al (2013) showed that the wealthy population is significantly more likely to share messages and photos on social platforms. However, we note a methodological challenge in studying the association between the peoples' socioeconomics conditions and the representativeness of VGI data, especially the representativeness of population flow obtained from VGI. There are mainly two methodological challenges, including high collinearity among demographic/socioeconomic factors and how to handle the spatial autocorrelation structure in flows. They pose high demands for effective statistical models for such biases. Thus, my third study first addresses the methodological shortcomings mentioned above and further explores data representativeness in VGI flow data.

RQ3: How do various demographic/socioeconomic attributes relate to the representativeness of VGI flow data? Which attributes are most strongly associated?

I still used Twitter data in the third part of the research. Many studies have examined Twitter data's distribution representative biases (Li, Goodchild, and Xu 2013). However, the representative biases in terms of spatial interactions and population flows are still far from clear. Regression models were always used to analyze representative biases. But spatial autocorrelation in population flows and highly correlated local demographic and socioeconomic characteristics is often simultaneously present in the model, impeding a valid regression. This part of the study aims to address these methodological difficulties and then assess the relationship between representativeness in Twitter flow data and demographic/socioeconomic attributes using the method we designed.

1.3. Dissertation Synopsis

This dissertation includes three thematic parts, illustrated in Chapters 2, 3, and 4, respectively. The remainder of this dissertation is organized as follows.

Chapter 2: Comparing Mobility Patterns between Residents and Visitors Using Geo-tagged Social Media Data

I designed a comprehensive analytical framework to evaluate and analyze socio-spatial segregation using human mobility information obtained from VGI. It includes collecting VGI data, classifying VGI users into groups (e.g., visitors and residents), extracting mobility data, constructing and analyzing interaction/mobility networks, evaluating dynamic socio-spatial segregation, representativeness analysis of VGI, and results' visualization. This part of the dissertation mainly focuses on analyzing mobility patterns for different VGI user groups, evaluating dynamic socio-spatial segregation, and representativeness analysis of VGI. The designed framework also serves as the basis for studies in Chapters 3 and 4.

Analyzing people's mobility patterns is the basis of studying their socio-spatial segregations. Researches have shown that different groups of people have different activity space and mobility patterns (Schönfelder and Axhausen 2003; Wang and Zhou 2017). Current studies mainly focused on the visitors' and residents' mobility patterns (Gabrielli et al. 2015; Orsi and Geneletti 2013; Li, Zhou, and Wang 2018). However, most of them considered all visitors as one group while overlooking the difference in mobility patterns between them. This chapter examined the activity spaces and movement patterns of different VGI user groups using the designed analytical framework. Specifically, I analyzed the mobility pattern of local Twitter users and visitor Twitter users based on the flow network and evenness distribution of user

activities. This study's results can provide decision-making information for tourism management, urban planning, and local economic development.

Chapter 3: an Extended Spatiotemporal Exposure Index for Urban Racial Segregation

This chapter refined the analytical approach in the designed analytical framework in Chapter 2. I designed and implemented the methodology to integrate new VGI data into the conventional segregation index. The Segregation Index quantifies the degree of segregation of social groups or classes. It provides practitioners a summary measure of the segregation within one area. However, with the rise of fine-grained spatiotemporal activity and flow data, the conventional segregation measurements' inclusiveness is challenged. This chapter extended the spatial exposure index by adding population flow patterns. Thereby it can describe spatiotemporal segregation changes with population movement. Specifically, the population flow network, hierarchical structure, and time information are used in the new extended spatial exposure index.

In Chicago's demonstration case study, I first estimated interactions between areal units at the neighborhood level using the time-dependent Twitter Origin-destination (OD) flow matrices and their hierarchical structure information. Then I computed the new population composition of units based on their interactions with other units. I also estimated the extended spatiotemporal exposure index for different time slots. Finally, I systematically compared their differences with the conventional indices at global and local scales. We can know better how the exposure index is affected after adding population flow patterns based on comparative analysis.

Chapter 4: Representative Bias in Spatial Movements and Interactions among Geotagged Social Media Flows Using Spatial Partial Least Square Regression

This chapter focuses on the representative bias in spatial movements and interactions among VGI flows. Representativeness of VGI is concerned by many studies since it does not have a systematic collection design and is not generated or collected to address a specific question. Many studies have examined the distribution representative biases of VGI, but the representative biases in terms of spatial interactions and flows are still far from clear. Regression models are always used to analyze representative bias. But spatial autocorrelation in VGI flows and highly correlated local demographic and socioeconomic characteristics is often simultaneously present in the model, impeding a valid regression. To address these methodological difficulties, I designed a Spatial Partial Least Square Regression approach to evaluate the association of neighborhood demographic/socioeconomic characteristics and representative biases in spatial movements and interactions among geotagged social media flows. Besides, to verify the designed approach's feasibility, a case study of 77 Chicago neighborhoods was conducted using geotagged Twitter flow data and Chicago Travel Survey. It analyzed the effects of five demographic and socioeconomic attributes on representative biases in spatial movements from Twitter data.

Chapter 5: Conclusions and Future Work

This chapter concluded findings from the three research articles in Chapters 2-4. I discuss limitations in this research and provide an outlook for future segregation research in the geographic domain.

CHAPTER 2: COMPARING MOBILITY PATTERNS BETWEEN RESIDENTS AND VISITORS USING GEO-TAGGED SOCIAL MEDIA DATA[1]

2.1. Introduction

Understanding residents' and visitors' behavior is vital in tourism studies and urban planning (Bauder and Freytag 2015; Dharmowijoyo, Susilo, and Karlström 2014). Investigating the mobility difference between residents and visitors can help urban planners and policymakers to improve the local economy and optimize the urban design. An influx of visitors into one area have both positive impacts (e.g., economic benefits, cultural diversity, and employment opportunities) and negative impacts (e.g., traffic congestion, competitive usage of public facilities, and damages to the built environment) (Weaver, Kwek, and Wang 2017; Andriotis and Vaughan 2003; Andereck et al. 2005). Understanding where and when the visitors appear in the city and their interactions with residents could help local agencies counteract visitors' negative impacts. Many studies have focused on the visitors' mobility pattern (Gabrielli et al. 2015; Orsi and Geneletti 2013; D. Li, Zhou, and Wang 2018)**,** without considering the heterogeneity of visitors, such as the variety of travel distance, the difference in culture, diversity of economic or demographic status, and distinction of travel behaviors (Vu et al. 2015; Batra 2009). Besides,

[1] This chapter is based on Liu, Qingsong, Zheye Wang, and Xinyue Ye. 2018. "Comparing Mobility Patterns between Residents and Visitors Using Geo-tagged Social Media Data." Transactions in GIS 22 (6): 1372–89. https://doi.org/10.1111/tgis.12478.

using traditional survey data in the study could be expensive and quickly become outdated. Therefore, there is a need to identify different groups of visitors and to depict their mobility patterns using geotagged big data and the corresponding analytics.

Over the past decade, VGI has been increasingly used to explore the locals' and visitors' mobility patterns from different aspects (Miah et al. 2017; Chua et al. 2016; Zhou, Xu, and Kimmons 2015). Studies found that locals' mobility patterns are related to the urban spatial structure (Y. Chen et al. 2017; Naess 2000; Kang et al. 2012). However, the relationship between the visitors' mobility patterns and the urban spatial structure remains unclear and needs to be investigated further.

This chapter aims to uncover different groups of visitors' mobility patterns and the difference in mobility patterns between locals and groups of visitors. I first classify social media users into four groups: local users, state visitor users, national visitor users, and international visitor users. I then analyze the mobility pattern for each group of users and compare the spatial structure of mobility patterns between locals and three groups of visitors. Finally, this study examines and compares activity space and spatial structure of mobility patterns for these four groups of users.

2.1.1 Volunteered geographic information and tourists' mobility studies

VGI offers a great opportunity to analyze human movements at a finer geographic and temporal scale, which traditional survey approaches cannot achieve. Geotagged social media data, such as text and photos from Twitter and Instagram, has been utilized extensively in the study of human mobility (Shaw, Tsou, and Ye 2016; J. a. Long and Nelson 2012). Social media platforms provide much information to their users, such as the most attractive spots and best

restaurants. The users are not only consumers of this information but also its producers. People like to plan or rate their travels or tours based on social media information (Leung et al. 2013; Amaral, Tiago, and Tiago 2114; Xiang and Gretzel 2010). Moreover, people increasingly rely on social media to share their locations, precious moments, photos taken during the tour, and their comments after the tour (Y. Hu et al. 2015; Y. Liu et al. 2014). Therefore, social media users have created a large volume of VGI, and researchers can use these VGI to study their mobility patterns (Zeng and Gerritsen 2014; Shoval and Isaacson 2007; Ferrari et al. 2011).

To analyze locals' and tourists' mobility patterns (visitors) in a specific city, we first need to determine their origin locations. Three types of methods have been used to infer users' origin locations in previous studies. The first type of method takes users' profile locations as their origin. Self-reported questions like "which city do you live in" or the time zone information generated by the social media platform could help us determine users' origin (D. Li, Zhou, and Wang 2018; Chua et al. 2016). However, the self-reported information could be inaccurate or incomplete. The second type of method uses activity information, such as the location of tweets or Instagram photos, to infer users' origin. More specifically, it takes the location with majority of users' activities as their origin (García-Palomares, Gutiérrez, and Mínguez 2015; Huang and Wong 2016; Luo et al. 2016). However, this method may provide incorrect origin location information because visitors may take fewer photos in their familiar cities than tourists' attractions (D. Li, Zhou, and Wang 2018). The third type of method uses machine learning algorithms to determine the origin places of users. For example, Girardin (2008) proposed a multivariate logistic regression model to infer users' origins**.** This model incorporates time spans, the density of photos in one area, and the number of cities. Han et al. (2018) used the deep learning method to classify tourists by their travel purposes. However, human behaviors are very

complex and can be affected by many factors; we need to find some way to verify the classification results obtained by these three types of methods. Besides, the classification quality is affected by the characteristics of the VGI dataset used in each method. For example, most of Flickr's photos are highlights of a tour (Girardin et al. 2008). In contrast, Twitter users often tweet about the mundane aspects of life, such as what they had for breakfast and friends they met.

Based on the origin place of users, we can classify them into different groups. After the classification of users, researchers can conduct further studies of their mobility patterns. There are two strands of research on mobility patterns. One strand focuses on human mobility itself. Researchers assume that visitors' characteristics affect their mobility pattern (Amini et al. 2014; Vu et al. 2015; Batra 2009). Here, the characteristics refer to socioeconomic status, demographic features, and place characteristics. In particular, place characteristics refer to land use and urban spatial structure of places such as the central business district, residential area, or rich/poor community (Amini et al. 2014). Researchers can use these characteristics of visitors to classify them into different groups or infer their travel behaviors. Vu et al. (2015) analyzed international visitors' behaviors to help manage tourism in Hong Kong. Huang and Wong (2016) studied the travel behaviors of rich and poor local Twitter users. They found that Twitter users from poor areas tended to travel longer distances than Twitter users from affluent neighborhoods. The other strand, on the contrary, uses people's mobility patterns to infer the characteristics of one place or one group of the population. Many researchers found that land use or built environment has a relationship with people's travel behaviors (X. Liu et al. 2016; Zheng et al. 2014; Pei et al. 2014). Therefore, it is essential to reconstruct connectivity between regions and infer the land use and regional functions, based on people's travel behaviors (X. Liu et al. 2016; Pei et al. 2014;

Frias-Martinez and Frias-Martinez 2014). This chapter uses VGI to subdivide Twitter users into homogeneous groups and then analyze the spatial distribution of different groups of users' activities and find the difference between visitors and locals at the county level.

2.1.2 Heterogeneity of visitors

Different groups of visitors tend to have different travel behaviors (Vu et al. 2015; Batra 2009). While many researchers studied residents' and visitors' behavior, some researchers found that visitors can be further divided into sub-groups (Kerkvliet and Nowell 1999; W. J. Phillips and Jang 2010). We can consider both non-spatial factors and spatial factors in the process of subdividing visitors.

Non-spatial factors refer to socioeconomic status (SES) variables, such as gender, age, income, education, and race (Carter, S.M. 1998; M.-P. Kwan 2008). Non-spatial data is collected mainly through activity diaries or self-administered diaries (Chua et al. 2016). Researchers can distribute their questionnaires to specific groups of people that they want to investigate. For example, Kwan (2008) surveyed the experiences of Muslim women in the post-September 11 periods. The survey recorded their social status, their daily activities, and oral history. Based on the information obtained from this survey, Kwan (2008) analyzed and visualized those Muslim women's activity patterns.

Spatial factors are related to spatial locations. Travel distances of visitors are the main spatial factor (Ahn and McKercher 2015; Eagles et al. 2015; Fang Bao and Mckercher 2008). There are many ways to measure the characteristics of visitors' travel distances. Bao and Mckercher (2008) defined travel distance as the distance from visitors' home to their destination; they found that long-travel-visitors have distinct behavior patterns from short-travel-visitors.

Besides the physical travel distance, cognitive distance and cultural distance are also very important in studying human activity or mobility patterns (Ankomah, Crompton, and Baker 1996; Ahn and McKercher 2015). The cognitive distance and cultural distance are situational constraints formed by the social-psychological process. They reflect each visitor's social, cultural, and life experiences and can be treated as the cognitive representation of visitors' actual travel distance (Ahn and McKercher 2015; Ankomah, Crompton, and Baker 1996).

Currently, we know little about the behavior and mobility patterns of subgroups of visitors. Due to the lack of this knowledge, locals lack the opportunity to promote their potential visitors. Moreover, the local government's policy may waste resources because of the spatial mismatch of tourism facilities and the visitors. In contrast, locals can take the positive impact from the influx of visitors and avoid the negative impacts if they know where and when the visitors would like to visit their places or cities. Fortunately, the geo-tagged social media data contains both spatial and temporal information. It can provide more information about the travel behaviors of subgroups of visitors. Therefore, it is feasible and essential to use geo-tagged social media data to extract or analyze spatial factors that need to be considered in the process of subdividing visitors.

2.2. Study area and data

According to statistics, 17.6 million visitors generate 8.1 billion dollars in the tourism sector in the Cleveland area in 2015 (Glaser 2016). The tourism industry is a significant contributor to its economic vitality. The study area in this chapter is the Greater Cleveland area (Figure 2.1). It contains ten counties. The upper five counties (Cuyahoga, Geauga, Lake, Lorain,

and Medina) are core urban areas, and the lower five counties (Wayne, Summit, Portage, Stark, Carroll) are its peripheral areas.

Two Twitter datasets were collected using the Twitter application programming interface (APIs) Stream and Rest. We first harvested tweets from Oct. 1, 2015, to Feb. 28, 2016, from Twitter using the Streaming API. Then we collected all the tweets located in the study area and discard those falling outside of the study area. In total, we collected 119,773 geo-tagged tweets, which were referred to as the Stream Dataset. Since we plan to classify social media users based on their spatiotemporal characteristics, we would like to collect more tweets for each user to make the classification results as reliable as possible. Therefore, we used the REST API (statuses/user_timeline) to retrieve the most recent 3200 tweets for each user. Finally, we got 2,287,120 recent geo-tagged tweets for all users, referred to as Recent Dataset. Including more past tweets of each user in the Recent Dataset provides users' activity information outside the study area, making our classification of Twitter users possible. We first use the Recent Dataset to subdivide Twitter users, and then we use the Stream Dataset to analyze the mobility pattern of each group of Twitter users.

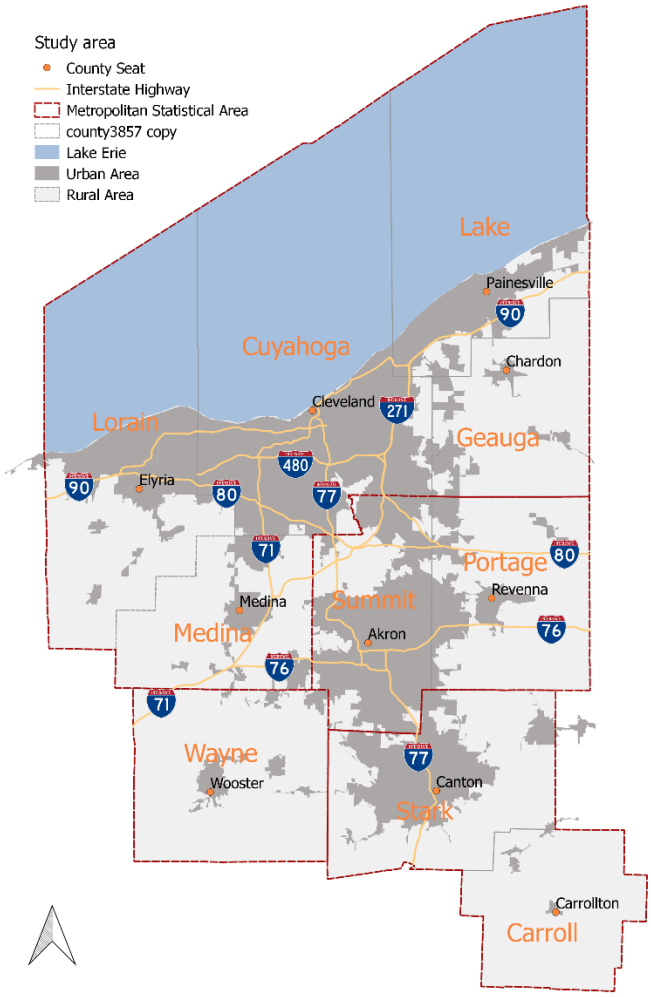


Figure 2.1 *Map of Greater Cleveland, OH*

2.3. Methodology

The research flowchart used in this study is shown in Figure 2.2. First, the Twitter users were classified into four groups according to their past activities, which were stored in the Recent Dataset (Figure 2.2a). The four groups refer to the Local Resident Group (Local Group), State Visitor Group (State Group), National Visitor Group (National Group), and International Visitor Group (International Group). After classifying Twitter users into these four groups, the candidate moves can be classified according to user group identification (Figure 2.2b). Finally,

we evaluate the activity space and the mobility pattern by calculating the evenness and the centrality indexes for each Twitter user group (Figure 2.2b). The detailed methodology is discussed in the following subsection.

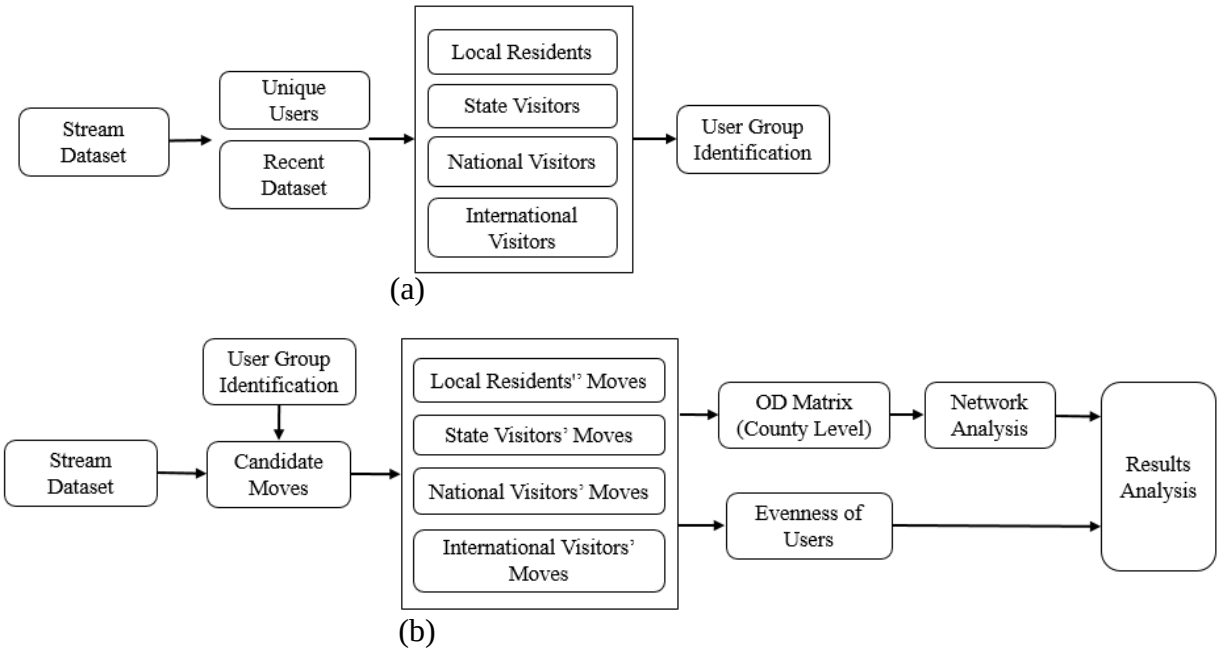


Figure 2.2 *Research Flow Chart: (a) Identify User Groups; (b) Main Analysis Flow Chart Combining the User Group Classification*

2.3.1 Classification of Twitter users

We classify users at four spatial scales (Local, State, National, and International). The geographic scales can be a proxy of the cognitive distance from users' home to the study area (i.e., Cleveland Metropolitan area). The spatial scale concept is widely used in daily conversations, especially to describe an unfamiliar place (e.g., West Side Market is one of the landmarks in Cleveland, Ohio). Therefore, it is reasonable to concatenate different spatial scale words to represent one place and classify users into different groups.

Based on the above spatial scales, each user's origin could be determined by the most frequently visited units of spatial scale from 8:00 pm to 8:00 am. The number of posting of each user was counted in each unit of spatial scale. As suggested by Huang (2016), we also set the cut-off value to be 40 when selecting eligible Twitter users (i.e., we only keep the Twitter users who posted more than 40 tweets). After the classification of Twitter users, we can get the user group identification in our dataset, indicating which group a Twitter user belongs to (Local, State, National, or International).

2.3.2 Mobility pattern of the four population groups

The mobility pattern of each group is represented by the Origin-Destination (OD) matrices, where each element represents the travel flow between two counties in the study area. We build a flow network based on the flows between counties and then use the flow network's centrality index to analyze the importance of each place in the flow networks for each group. Besides, we also compare the mobility pattern of locals with that of three groups of visitors. In the following subsections, we discuss our method to generate the OD matrix and calculate the centrality index of the flow network in detail.

2.3.2.1 Generate OD Matrices for the Four Population Groups

To build the OD matrices, we need to extract the travel flow between any two counties for each group of users. We organize the tweets from Stream Dataset into a temporally ordered sequence for each user. Let S denotes this sequence, and S has the form like$\{s_0 = (l_0, t_0, c_0), s_1 = (l_1, t_1, c_1), \dots, s_n = (l_n, t_n, c_n)\}$, where l_i represents the geo-location (latitude, longitude) of the tweet s_i, with contents c_i at time t_i. Then, each two consecutive temporal tweets (s_i, s_{i+1}) of one Twitter user, can denote this user's one **move** from location l_i to location

l_{i+1}. However, not all moves meet our requirements, so we apply two filters on these moves. First, we filter out the moves with a long temporal gap (i.e., $t_{i+1} - t_i > 4$) (Gao et al. 2014), because there may exist missing locations during a long temporal span (i.e., 4 hours) (Straumann, Çöltekin, and Andrienko 2014). Second, we filter out the moves with small distance shift (i.e., the distance between the l_{i+1} and l_i is less than 100 m), which may be caused by the uncertain GPS signal of a smartphone or the random walking around the same place. After the two filtering steps, we get all the ***qualified moves***. We use the "point in polygon" procedure in GIS to determine each move's origin county and destination county based on the qualified moves. Finally, we aggregate their qualified moves with the same origin and destination county to get entries in the OD matrix for each user group. Each entry in the OD matrix represents the total number of qualified moves between two specified counties. Since we have four groups of users (local, state, national, and international), we have four OD matrices, with a dimension of 10×10.

2.3.2.2 Centrality index of a flow network

The centrality index identifies the importance of nodes in a flow network graph. Here, we consider each OD matrix $A_k(k = local, state, national, international)$ as a flow network graph in which counties are nodes and travel flow between counties $a_{i,j}$, i.e., the travel volume from county *i* to county *j*, are edges. In the current literature, computer scientists and geographers have developed a number of indices for evaluating centrality in different dimensions (Crucitti, Latora, and Porta 2006). In this study, we choose one of the centrality indices used by Hughes (1993) for demonstration purposes because we have the same study scenario. Besides, this index includes a standardization procedure, and with this procedure, we can compare centrality results between our four user groups. Hughes's (1993) equation is shown in Equation (2.1).

$$c_i(\alpha, \beta) = \sum_j (\alpha + \beta c_j) a_{ij} \tag{2.1}$$

$$\sum_i c_i(\alpha, \beta)^2 = 10 \tag{2.2}$$

where α is a scaling vector, it is set to normalize the measurement; β is the range of interactions being considered (β= 0 in this study, because the travel flow is a direct flow from origin to destination for each group of users) (Hughes 1993); c_j is the centrality measurement of county *j*; and a_{ij} is the travel flow from county *i* to county *j*. In order to make our results comparable with results of other studies and make it reasonable to compare centrality indexes between different groups of users, we use Equation (2.2) to select the appropriate α value. Using this α value, we can get a normalized centrality measurement. Centrality index can represent the importance of one node in the network. If the centrality index of a node in the flow network is high and greater than 1.0, the corresponding node is in a central or important position in the network. And if the centrality index is less than 1.0, the corresponding node is in a peripheral position in the network. By considering the centrality level of ten counties for each group of users, we can understand the preference of activity space for each user group.

2.3.3 Spatial distribution of evenness

Users' sharing behavior on social media platforms varies greatly, with some users tweeting far more than others. To mitigate the bias caused by overactive users, we decided to use the evenness maps of Twitter users to show the spatial distribution of each group's activity, rather than using density maps. First, we divide the study area into hexagons of a specific size. Then for each hexagon in the study area, we calculate its evenness index, representing the

evenness of Twitter users in one hexagon. After calculating the evenness index of all hexagons in the study area, we can get the spatial distribution of Twitter users' evenness across the study area.

The evenness index used in the study is adopted from Shannon Entropy (Batty 2010). The user's evenness in one hexagon depends on the number of unique users (the richness of users) and the number of tweets posted by each user in the hexagon. The evenness index can be calculated using:

$$H = -\sum_{i=1}^{m} p_i log_2(p_i) = -\sum_{i=1}^{m} \frac{n_i}{N} log_2\left(\frac{n_i}{N}\right) \quad (2.3)$$

where *H* is the Shannon Entropy index, *m* indicates the number of users in one group, p_i is the ratio of the number of tweets posted by user *i* (n_i) to the total number of tweets posted in the hexagon (*N*), $\sum_{i=1}^{m} p_i = 1$. A high H value indicates high flows of people and information to or from that hexagon and vice versa. For a given number of coming users in one hexagon (*m*), the Shannon Entropy index reached a maximum when each user posted the same number of tweets, which is $p_i = 1/m$.

2.4. Results

2.4.1 Results of exploratory data analysis

We filtered out 155,844 qualified moves for 3,612 users. Table 2.1 shows the number of users, qualified moves, intra-county moves, and inter-county moves for each group. From Table 2.1, we can see local users account for 93.3% of the total qualified moves. In contrast, the visitor users (State, National, and International) only account for 6.7%, which is far less than local users. The higher percentage of local users is consistent with our expectations since the visitor

users only stay for a short period. The intra-county moves are always the dominant type of movement among their candidate moves for all groups of users. For example, the intra-county moves of local users are seven times more than their inter-county moves. In other words, the travel demand inside the county is much higher than that outside the county for all groups of users. For the state visitor users, 24.5% of its qualified moves are inter-county moves. This percentage is much higher than that of other groups (12.2%, 17.7%, and 10.6% respectively for local users, national users, and international users). The higher percentage of state users may reflect that state users have more connections with locals than other groups of users, considering its high ratio of inter-county moves.

Table 2.1 *Individual Flow Data at Each Level*

GROUPS	**# USERS**	**# QUALIFIED MOVES**	**# QUALIFIED MOVES**	
			# intra-county moves	# inter-county moves
LOCAL	2,019 (56%)	**145,401 (93.3%)**	127,600 (93.9%)	17,801 (89.5%)
STATE	370 (10%)	4,013 (2.6%)	3,030 (2.2%)	983 (4.9%)
NATIONAL	1,055 (29%)	5,961 (3.8%)	4,901 (3.6%)	1,060 (5.3%)
INTERNATIONAL	168 (5%)	469 (0.3%)	419 (0.3%)	50 (0.3%)
TOTAL	3,612 (100%)	155,844 (100%)	135,950 (100%)	19,894 (100%)

Figure 2.3 shows the visualization of the flow network of each group. The direction of blue and orange arrows indicates the direction of moves, and the width of the arrows is proportional to the volume of flow. The start and end point of the flow is the county seat location, represented as a circle in Figure 2.3. The size of the circle means the flow volume of the intra-county moves. The network in Figure 2.3a clearly shows that Cuyahoga County and

Summit County are cores of in and out flows, and the two counties comprise the most critical part of the whole flow network in the study area. The flow networks of state users and the national users share a similar pattern, in which the flow volume is evenly distributed in all directions. Therefore, the connectivity in the flow network is fully developed. The international flow network is shrinking to the core area, concentrating in Cuyahoga and Summit Counties.

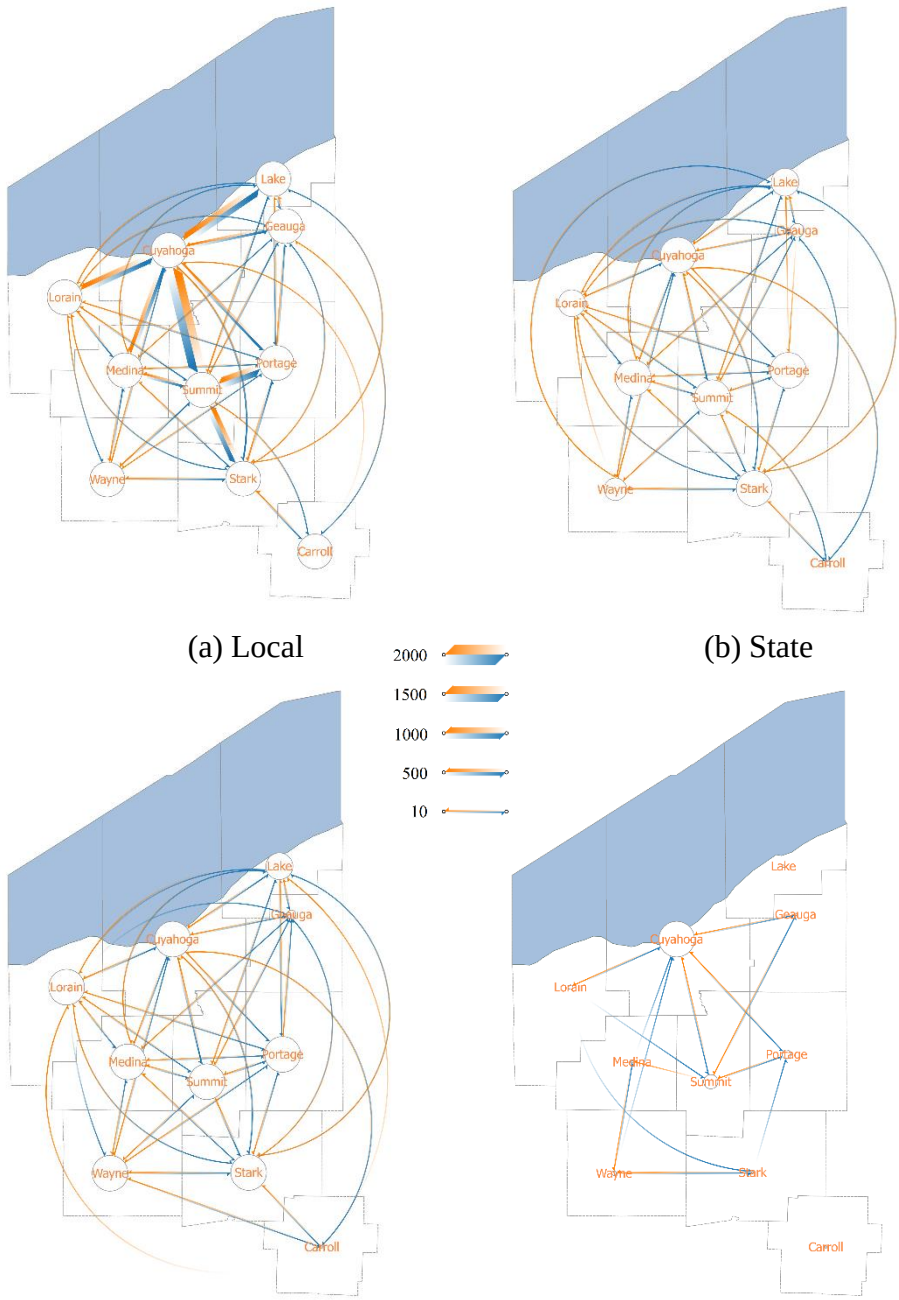


Figure 2.3 *User Flow Visualization: (a) Flow Network of Local Users; (b) Flow Network of State Users; (c) Flow Network of National Users; (d) Flow Network of International Users.*

4.2 Comparison of Mobility Patterns

Table 2.2 and Figure 2.4 show the centrality indexes of each group in each county. The order of counties in Table 2.2 and Figure 2.4 is in descending order of its population size. Cuyahoga County has the largest population, and Carroll County has the smallest. Through Table 2.2, we recognized the core-peripheral structure in the study area. First, the centrality index of Cuyahoga and Summit County is greater than 1.0, meaning these two counties are center places for visitors and locals. The recognized core area is consistent with our expectations. Since the 1990s, Cleveland has experienced an unexpected renaissance after the decline of industrialization, including an emerging high-tech sector, producer services, and a center for cultural consumption (Warf and Holly 1997). These centers are still developing around the core of the region. Second, Stark, Lorain, Lake, Medina, and Portage County compose the first peripheral layer around the central areas. They have the centrality index from 0.5 to 1.0. Third, Wayne, Geauga, and Carroll County have a centrality index lower than 0.3. They compose the outmost peripheral layer in the core-peripheral structure.

Table 2.2 *Centrality Index for Four Groups*

	Local	State	National	International
Cuyahoga	2.20	2.17	2.24	2.12
Summit	1.85	1.63	1.73	1.60
Stark	0.56	0.66	0.73	0.43
Lorain	0.45	0.66	0.53	0.59
Lake	0.59	0.61	0.52	**0.00**
Medina	0.50	**1.06**	0.56	0.52
Portage	**0.74**	0.39	**0.72**	**1.11**
Wayne	0.11	0.13	0.14	**0.67**
Geauga	0.26	0.32	0.24	0.29
Carroll	0.03	0.04	0.05	0.00

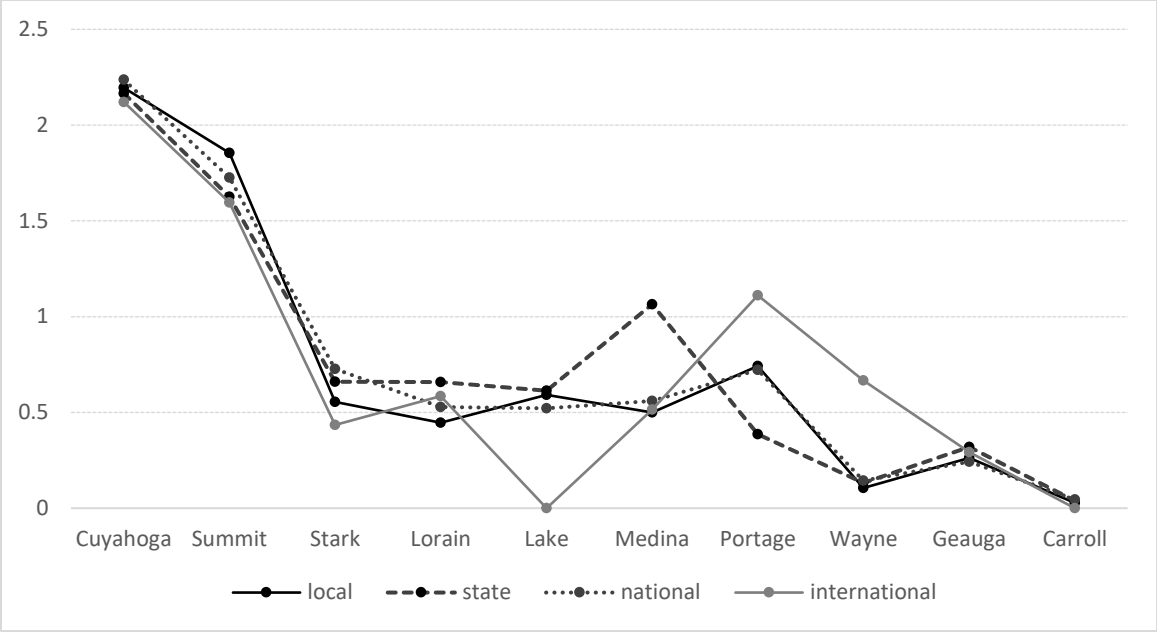


Figure 2.4 *Visualization of Centrality Index for Four Groups*

Figure 2.4 is a visualization of the centrality results of the four groups in Table 2.2. From Figure 2.4, we can see that the centrality index curve for each group generally tends to decrease with the decreasing of population size; however, there are some exceptions (bold in Table 2.2). For instance, Portage County has a smaller population than Medina County, but its index of centrality in the local user network is larger than that of Medina's. We cannot be sure why these exceptions occur, whether it is due to random factors (i.e., the uncertainty of the VGI flow data) or systematic factors (i.e., higher education schools in Portage). We may need more investigation and data collection to answer this question.

The centrality index of local users is more similar to that of national users than state users. In addition, differences in centrality indices between local and state users exist primarily in Medina and Portage counties. The centrality index in the national user flow network of Medina

County is much higher than the centrality index of the other three groups of users. But the situation in Portage County is the exact opposite. Through Figure 2.1, we can observe that the distance from both counties to the core area (Cuyahoga and Summit County) is comparable. But why do they show such a big difference in the flow network of the state user group? We suggest that the most likely reason for this difference is that state visitors are from other Ohio parts, and their collective flow network is influenced by the connections between the Cleveland area and other metropolitan areas. For example, Medina County is on Interstate Highway 71, which connects the Cleveland metro area with the Columbus metro area. Portage County is on Interstate Highway 80, which connects the Cleveland metropolitan area with the Youngstown-Warren metro area. The higher centrality index of state users in Media County and lower centrality index of state users in Portage County may be caused by the following reasons. 1) More visitors are coming to the study area from Columbus metropolitan area than from Youngstown-Warren metro area. 2) Cleveland metro area has a stronger connection to the Columbus metro area than to the Youngstown-Warren metro area. 3) There are more attraction sites along highway I-71 than highway I-80 in Ohio State. However, the complex connections between metropolitan areas and factors affecting people's mobility need further study.

2.4.2 Spatial Distribution of Evenness

A total of 311,688 tweets were assigned to the hexagon grids using a "point in polygon" operation in QGIS. The evenness index of each hexagon grid was calculated for all groups of users. As mentioned before, the evenness index can be affected by the hexagon's size; we calculate the evenness index at different spatial scales by setting the size of the hexagon to be 10, 5, 2.5, and 1 km, respectively. In other words, we calculated the evenness index at different

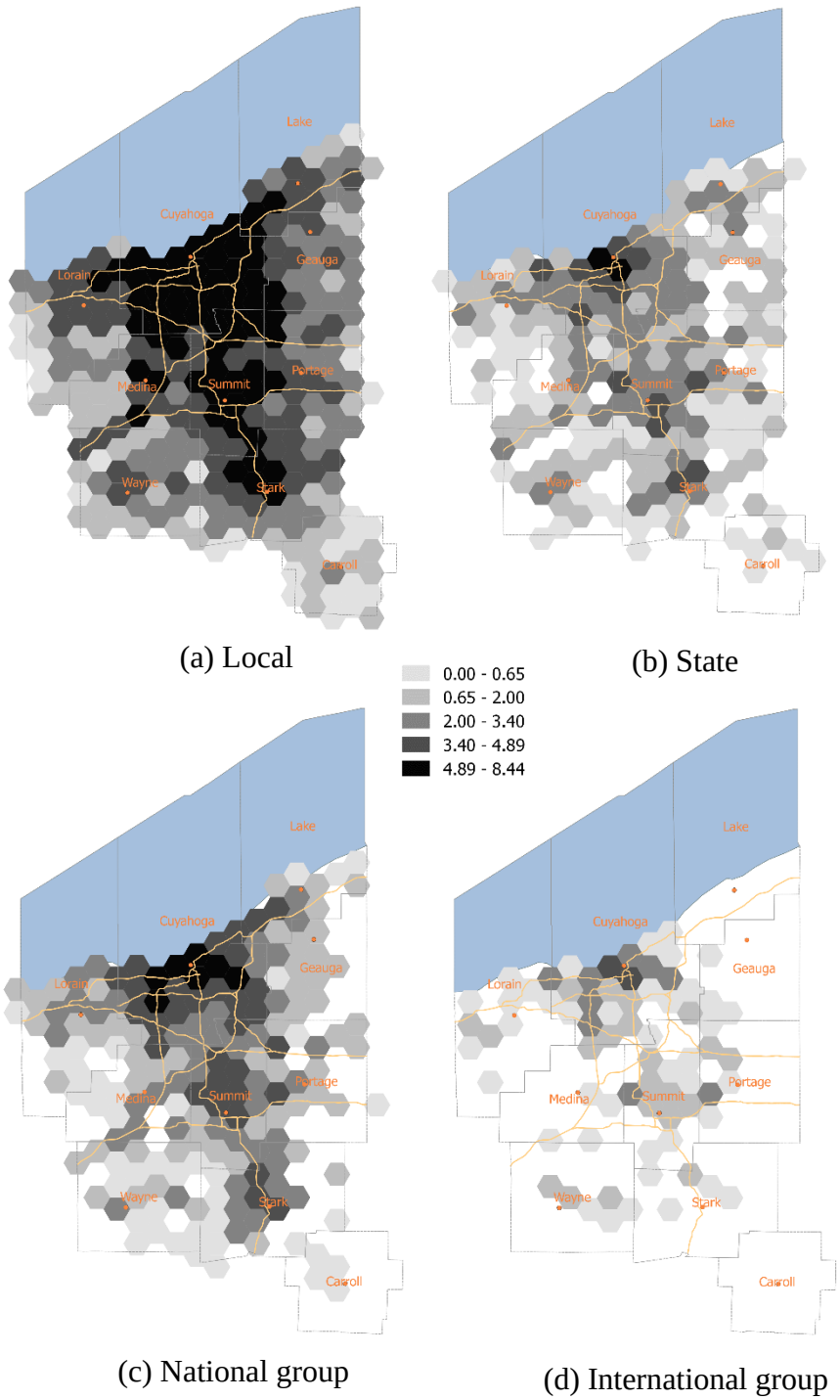
spatial scales in the study area. The results are shown in Figure 2.5-2.8. The darker the color, the higher the evenness index value. An area with a high evenness is more frequently visited by Twitter users. The yellow lines in Figure 2.5-2.8 are interstate highways in the study area; the orange dots represent the county seats.

From Figure 2.5a, we found that Local group users like to visit almost all the urban areas of Cuyahoga County, Summit County, and Stark County, which were colored black. If we define the darkest area as the activity spaces of Local group users, they are consistent with the Metropolitan statistical urban areas (dark grey color) in Figure 2.1. For the other three groups of users, their activity spaces were also concentrated in the urban area. However, their range tends to shrink compared to that of Local group users. For example, the most commonly visited places of international visitors (Figure 2.5d) are mainly located in Cleveland City, which much smaller than that of the other three group users.

From Figure 2.5a-d, we can see that among the four groups of users, the county seat of Cuyahoga County (Cleveland) is constantly the most commonly visited place by all users (in the darkest color). Moreover, these commonly visited places of residents form a "T" shape. When the distance from the "T"-shaped areas increases, the evenness index began to decrease. Twitter users in the "T"-shaped areas are more diverse than the outside areas. Therefore, we can say that the spatial distribution of Twitter users' evenness forms a generally concentric pattern, from the "T"-shaped area to peripheral areas. The concentric pattern is also consistent with our centrality analysis results in the previous subsection. Besides, all the "T"-shaped areas have interstate highways passing through them. It may indicate that the interstate highways have some impacts

on Twitter users' activities. However, we need more data to verify the impacts of highways on the users' evenness.

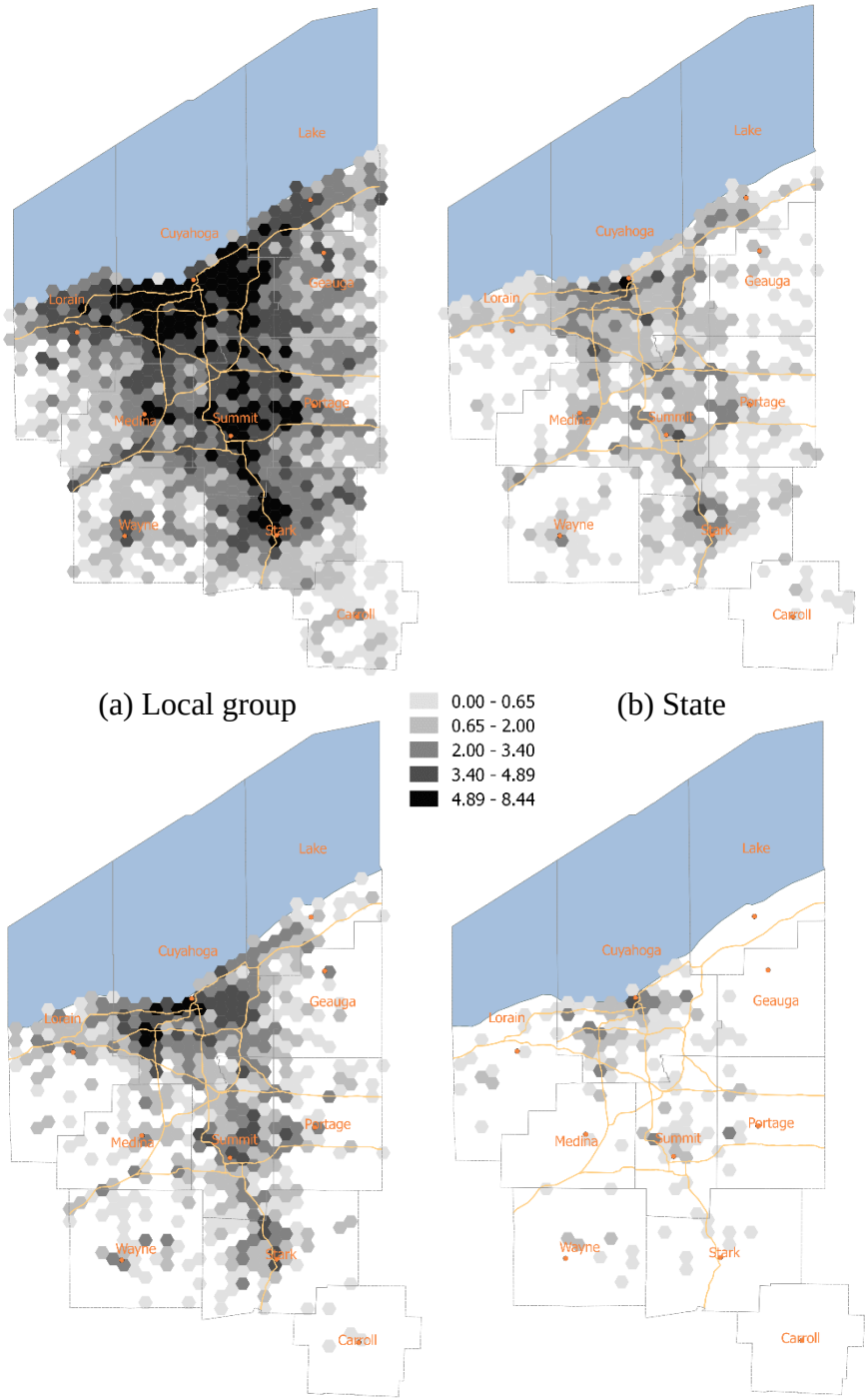
In Figure 2.5b, Cleveland is the most frequently visited place by state users. Moreover, the other frequently visited areas (dark grey) are also in each county's urban areas. There is also a "T"-shaped region in Figure 2.5b. National users have almost the same spatial distribution of evenness (i.e., Twitter users' activity patterns). All the places visited by national visitors are along the interstate highways. The concentration of activity around the highway may also reflect interstate highways' impacts on national visitors' activities. However, we need further study to see if the interstates can help attract more visitors to the area. Comparing the results in Figure 2.5-2.8, we can see that the spatial scales affect evenness distribution. With the decrease of hexagon size, the most commonly visited places shrink to smaller areas. As an example for local users, the black area shrinks from the entire urban area of Cuyahoga in Figure 2.5a to the spotted area near Cleveland in Figure 2.8a. However, the "T"-shaped patterns are retained across different spatial scales.



(a) Local (b) State

(c) National group (d) International group

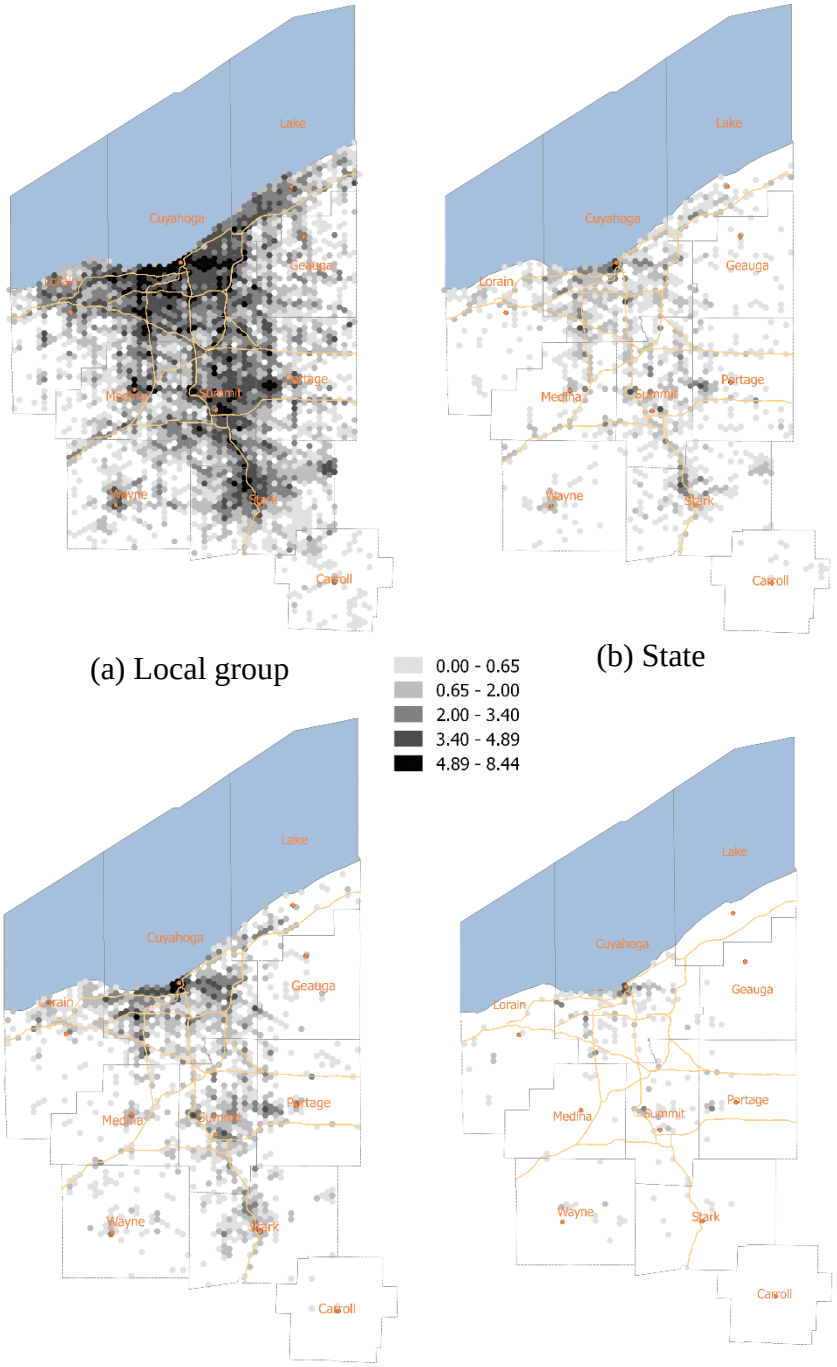
Figure 2.5 *Evenness Distribution for Four Groups at the Spatial Scale of 10km Hexagon: (a) Evenness Index for Local Residents Group; (b) Evenness Index for State Visitors Group; (c) Evenness Index for National Visitors Group; (d) Evenness Index for International Visitors Group*



(a) Local group (b) State

(c) National group (d) International group

Figure 2.6 *Evenness Distribution for Four Groups at the Spatial Scale of 5km Hexagon: (a) Evenness Index for Local Residents Group; (b) Evenness Index for State Visitors Group; (c) Evenness Index for National Visitors Group; (d) Evenness Index for International Visitors Group*



(a) Local group (b) State

(c) National group (d) International

Figure 2.7 *Evenness Distribution for Four Groups at the Spatial Scale of 2.5km Hexagon: (a) Evenness Index for Local Residents Group; (b) Evenness Index for State Visitors Group; (c) Evenness Index for National Visitors Group; (d) Evenness Index for International Visitors Group*

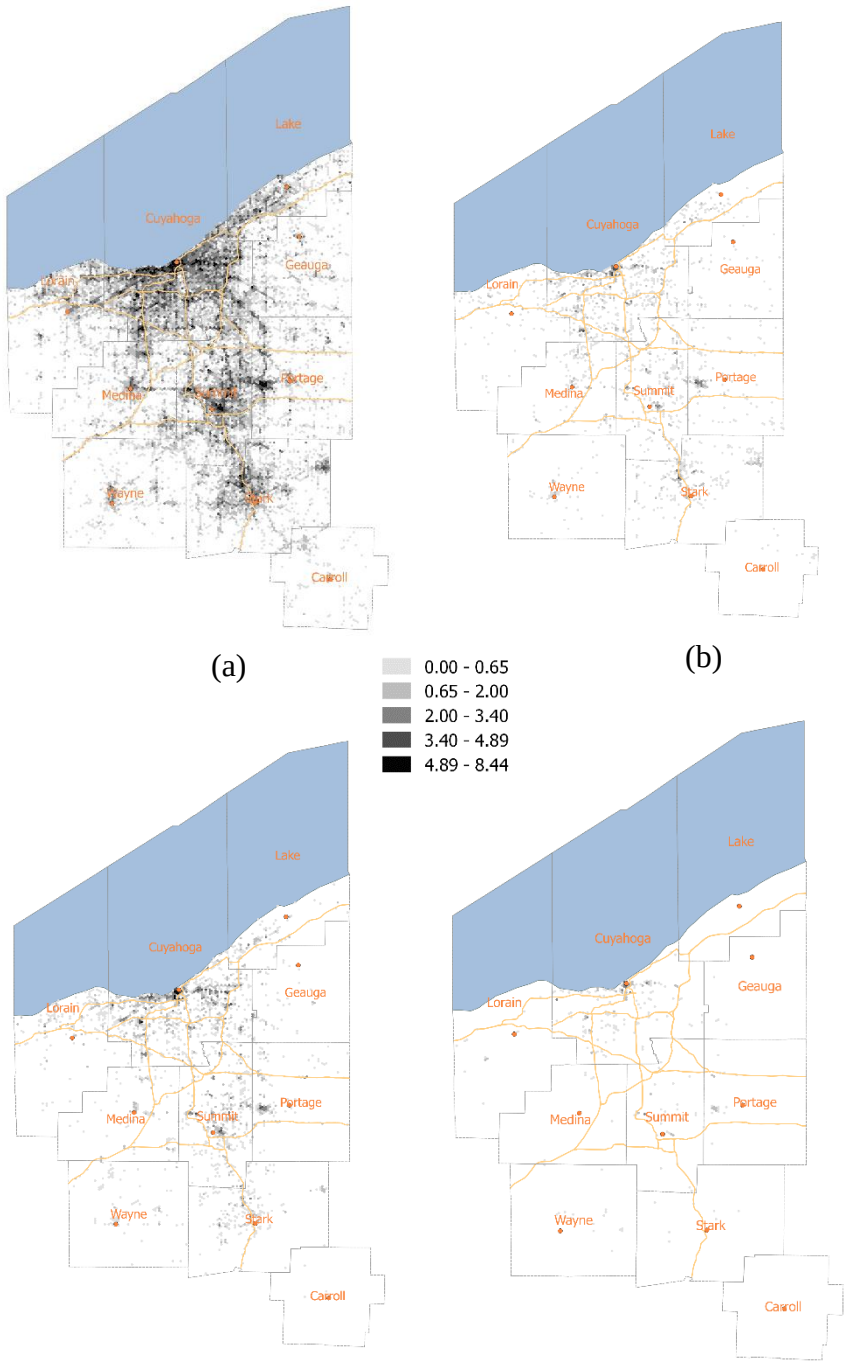


Figure 2.8 *Evenness Distribution for Four Groups at the Spatial Scale of 1km Hexagon: (a) Evenness Index for Local Residents Group; (b) Evenness Index for State Visitors Group; (c) Evenness Index for National Visitors Group; (d) Evenness Index for International Visitors Group*

2.5. Discussion and Conclusion

Geotagged social media data is increasingly used in tourism study and urban planning. However, the difference in mobility patterns between sub-groups of visitors and locals has not been fully investigated. This chapter analyzed the mobility patterns of local and visitor Twitter users with flow networks and evenness distribution of Twitter user activities. First, we explored the basic mobility pattern of local Twitter users and visitor twitter users (including state users, national users, and international users). Table 2.1 and Figure 2.3 found that short distance movement is the dominant type of activity for locals and visitors. Moreover, intra-county movement accounts for the main type of movement for all groups of Twitter users.

Second, we reconstruct the core-peripheral structure in the study area based on Twitter users' centrality index for the four groups. Cuyahoga and Summit County are the core; Stark, Lorain, Lake, Medina, and Portage County compose the first peripheral layer around the central areas; Wayne, Geauga, and Carroll County form the outmost peripheral layer. Based on the centrality index alone, the visitors' flow network's characteristics follow the same general trend as local users, with minor differences comparing the four groups. Moreover, the centrality index of local users is more similar to the one of national users than to one of the state users. The difference between users becomes larger in counties at the first peripheral layer. From the analysis, we found that the flow pattern of locals and visitors could be affected by population size and its connections with other regions outside the study area. However, further study is needed.

Third, from the spatial distribution of the evenness index at different spatial scales, we found that the Downtown area in Cuyahoga County is commonly visited by both locals and

visitors. A clear “T”-shaped core-peripheral structure is observed from the evenness index maps in Figure 2.5-2.8. Besides, we suspect that the distribution of tweets is correlated with the direction of the interstate highway.

Even though we gathered some initial insights and information from VGI in this study, there are still some limitations. First, current VGI, especially from social media data, has several biases regarding place accuracy and demographic composition. For example, geo-tagged tweet data can only reflect the mobility characteristics of a particular population. According to Aslam (2017), about 66% of Twitter users are between 18 to 49 years old, and 54% of Twitter users earn more than $50,000 a year. Thus, the tweets mobility data cannot fully depict the mobility of the older and poorer population groups. Second, we only use tweet data. However, we should also be aware of other available open data, such as the authority survey data from U.S. Census Bureau (e.g., American Community Survey) and other social media data (e.g., Foursquare, Yellow page, Instagram, and TripAdvisor, etc.). These data sources also provide rich information about people’s mobility patterns from different aspects. For example, Instagram provides us with geo-tagged photos from visitors, TripAdvisor provides information about points of interest, including their locations and comments from local and visitor users. We hope to synthesize multiple open big data to improve the study in this chapter. Furthermore, all the analysis in this study is at the county level, giving us information at large (i.e., coarse) spatial scale. However, to get more detailed information about human mobility patterns and find the factors affecting human activities, we need to conduct our study at finer scales, such as census tracts, census blocks, or even street levels.

CHAPTER 3: AN EXTENDED SPATIOTEMPORAL EXPOSURE INDEX FOR URBAN RACIAL SEGREGATION[2]

3.1 Introduction

The spatiotemporal activity and trajectory data play an increasingly important role in measuring the degree of segregation of social groups or classes (Wong and Shaw 2011; Yip, Forrest, and Xian 2016; Farber et al. 2015). Socio-spatial segregation implies a lack of communication between groups and indicates an uneven distribution of population or resources and varying interactions. Due to data limitations, the conventional segregation indices rely heavily on census data and focus on static residential spaces. Researchers have argued for expanding the analytical focus to other relevant places and times in people's everyday lives (M.-P. Kwan 2015). Over the past decade, spatiotemporal activity and trajectory data helped researchers to characterize changes in individual or group activity spaces (D. Wang, Li, and Chai 2012) and flow patterns (Guo et al. 2012; Gao et al. 2013) at different time scales (Silm and Ahas 2014b). The observed people's movement connects the city's various spaces and creates an entire network of interactions. In turn, the population flow network and its interaction information are instrumental to our understanding of cities (Batty 2013).

[2] This chapter is submitted to the journal of Cartography and Geographic Information Science for peer review.

However, flow patterns, which contain interaction information, have not been fully tapped in segregation studies. The first question focuses on the methodology: how can we incorporate flow patterns into traditional segregation calculations with minimal modifications of conventional formulas? The second is observing indices from a comparative perspective: when comparing the new index to the conventional one, where and when do the two exhibit significant differences? What factors influence the extent of their differences? Answering the above questions will not only allow us to build a comparable system to link traditional research and current research with new data, to form a succession and continuation of research, but also to provide a baseline for using the richer spatiotemporal activity and flow data.

This chapter attempts to incorporate time-dependent population flow patterns into conventional segregation indices computation. Specifically, we use the flow network information, time, and hierarchical structure to estimate interactions between areal units. By utilizing more information, we expect to comprehensively evaluate the interactions and then compute the segregation more accurately. Besides, this chapter also systematically compares the new flow-based segregation index with the conventional indices. Based on the comparison results, we try to understand better the impact of including population flow patterns into the segregation indices.

This chapter is organized in the following way. Section 3.2 briefly reviews the development of the segregation index and the progress of methods using new data. Section 3.3 covers the study area and demonstration data. Section 3.4 illustrates the methodology used in this chapter, including how to construct flow patterns, compute the segregation with flow data, and the comparison method. Section 3.5 shows the results. We begin with a descriptive analysis of

flow patterns by different neighborhood types in the Chicago area and then compare the segregation index changes after incorporating flow patterns at both the global and local scales. In Sections 3.6, we discussed the implications of adding flow data to segregated studies and insights from the comparative results. Section 3.7 gives a brief conclusion.

3.2 Literature review

The spatial and temporal perspective provided by geographers is far-reaching when discussing how to quantify segregation. Since the dissimilarity index (Duncan and Duncan 1955), many indices have been proposed, and vast literature has continually criticized, improved, and generalized these segregation indices. Their basic principle considers areal units' spatial arrangement/structure in the study area when evaluating segregations. Spatial dependence or interactions is the most discussed spatial structure in the literature. Without considering spatial interactions, the corresponding segregation degree depends only on the areal unit's demographic composition. It raises the checkerboard problem, an essential methodological issue (Morril 1991; Wong 1993). How to accurately describe these interactions between units becomes the key to solving the checkerboard problem.

There are two main classes of methods to describe these interactions. One is based on spatial proximity functions, and the other is based on topology or geometry information. The spatial proximity functions follow the principle of distance decay and assume that the farther the distance between two areal units, the weaker their interactions. White (1983) suggested four proximity functions, including negative exponential functions and inverse distance functions with different parameter settings. With an empirical comparison between these proximity functions, White concluded that the issue of choosing appropriate proximity function forms and

parameters is still not yet resolved. Given that there is no better way to select a spatial proximity function and its related parameters, Reardon & O'Sullivan (2004) left the choice to users. While generalizing his predecessor's work, he used a notation to represent this spatial proximity function in terms of decreasing distance. Reardon et al. also emphasized that any desired proximity function could be used in his proposed generalized framework. Reardon's framework is promising but left subsequent researchers with the difficult task of quantifying the spatial proximity function.

The second class of methods uses topology or geometry information to describe spatial interactions between areal units. A simple topology-based approach is to use the binary form of contiguity. If areal unit i and j are neighbors, their corresponding weight, $\omega(i,j)$, is set to 1, and 0 otherwise (Morril 1991). Subsequently, Wong (1993) incorporates geometry properties, such as shape and area of units, to adjust the interactions. Wong assumes that longer shared boundaries would lead to a stronger interaction between two units. Wong (1998) also proposed a notion of composite population count (CPC), which mixing its neighbors' populations into the target unit population. After the mixing procedure, the new population composition is used in the subsequent calculations. In essence, CPC is a way to describe the probability of the target unit population interacts with its neighbors' population. In short, both above classes of methods attempt to provide descriptions of neighbor interactions and eliminate the checkerboard problem.

Although these two estimation methods can solve the checkerboard problem, their estimation of the interactions is subjective. No matter how complex the mathematical forms, there is no guarantee that their interaction estimation reflects the actual social interaction across units. With the development of information technology, the accumulation of massive amounts of

spatiotemporal activity and trajectory data provide new insights into modeling or estimating social interactions. First, researchers emphasize expanding static residential space into other socio-geographical spaces (Wong and Shaw 2011; Yip, Forrest, and Xian 2016; M.-P. Kwan 2015). Many studies have demonstrated, from different perspectives, that new data can provide additional information than traditional survey data. For example, Shelton et al. (2015) examined residents' tweeting activity on both sides of the imaginary "9th Street Divide" from collective activity distribution perspective. Q. Wang et al. (2018) analyze geotagged tweets to compare the travel patterns for 50 U.S. cities by racial/ethnic groups. Park and Kwan (2018) proposed the individual segregation index to depict when and how much segregation people experience dynamically throughout a day. Second, researchers emphasize the importance of the interaction between physical and social space in the study of segregation. Farber (2015) resorted to spatiotemporal paths to describe individual trajectories and further assessed interaction-based segregation. Xu (2019) used the call detailed records (CDR) to portray interpersonal communication intensity and then estimated socio-spatial segregation based on friendship networks. These examples show the promise of using new data to describe the distribution of activities or interactions between regions within cities from different perspectives.

However, current conventional segregation studies do not fully explore the dynamic interaction information from population flow patterns. First, without changing the conventional segregation index formula, population flow networks have not been fully discussed. We argue that the population flow network can be easily integrated into conventional segregation indices with only a few uncomplicated transformations. The linkage exists in the notions of the composite population count (CPC) (Wong 1998) or population density of the local environment (Reardon and O'Sullivan 2004) or local population intensity (Feitosa et al. 2007). All notions

emphasize the exchanges/interactions between local residents and their neighbors, and thus, measuring the intensity of such interaction becomes the key for segregation computation. In the CPC notion, the interactions between two areal units are estimated based on their contiguity. In comparison, the other two notions underscore the distance decay functions in the estimation. Although researchers have made many modifications to the contiguity or set up many distance decay functions, the assessments are still artificial.

Second, collective-level interactions change with time, and the temporal dynamics need more attention. The traditional definition of interaction is not only fixed and artificial in the spatial dimension but also the temporal dimension. For example, if using contiguity to construct spatial interactions, one's neighbors do not change from ten years ago to ten years later. The same problem exists for the method of distance decay functions. As the data becomes more abundant, researchers have realized the necessity of focusing on temporal dynamics in segregation studies (Silm and Ahas 2014b; J. Lee and Li 2017; M.-P. Kwan 2015); however, the temporal variations in interactions between areal units have received little attention. Equally as important as the spatial dimension, the temporal interaction variations tell us when interracial contact is lower or higher. This temporal pattern can help us better understand segregations in today's fast-paced and mobile world and provide a basis for reviewing integration and planning policies (Silm and Ahas 2014b). By adding population flow information to estimate interactions, one's neighbors can change with commuters traveling to work in the short term and change with the city's development in the long run. Therefore, using population flows to determine spatial interactions can avoid the static problem of segregation index.

Finally, the hierarchical structure of population flow has received little attention in segregation studies. Due to information technology and transportation accessibility, residential segregation becomes less and less important, and contact with distant neighborhoods or central cities shapes the external connections of the local residents. While local neighborhoods may be segregated, residents do not live only in their neighborhoods. We should pay more attention to the mobility of residents and their position in social and physical networks (Browning and Soller 2014). And the hierarchical structure consisting of mobility networks in the city is one way to evaluate the importance of different neighborhoods. Bassolas et al. (2019) used flow data to assess the hierarchical structures of 301 cities worldwide and identified different hotspot levels for areas in each city. The study found that cities with a more robust flow hierarchy have a higher degree of population-mixing, extensive public transportation, and a higher walkability level. Liu et al. (2018) found that the core-peripheral flow structure is in accordance with the visitors' distribution in one area and affects the residents' and visitors' meeting probability. The hierarchical information facilitates a comprehensive picture of segregation within the urban's current spatial layout. Thus, the segregation indices with hierarchical structure information can be a meaningful reference for urban planners.

3.3 Study Area and Data

The 77 Chicago neighborhoods (Chicago Data Portal n.d.) have three main racial/ethnic groups, including approximately 29% non-Hispanic Black, 32.5% non-Hispanic White, and 29% Hispanic. Based on the population information of 2012-2016 American Community Survey 5-year Estimates, Figure 3.1 illustrates where the three groups of the population living in Chicago. On the map, the colors of aqua, red, and green dots represent non-Hispanic Black, non-Hispanic White, and Hispanic groups, respectively (100 persons per dot). The non-Hispanic Black

population group is predominantly located in the South and West. In contrast, the non-Hispanic White population group is primarily located in the North, and the Hispanic population group is interspersed between the other two groups.

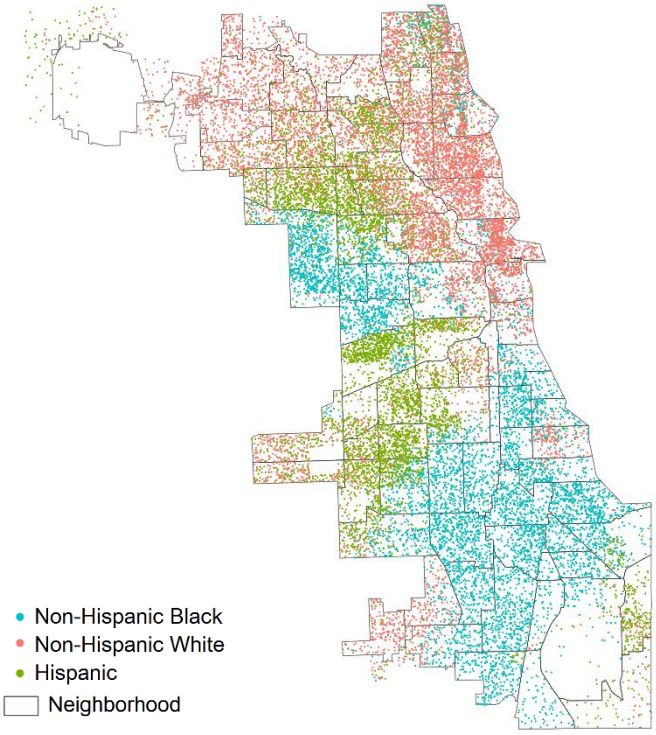


Figure 3.1 *The Population Distribution of Three Groups within Chicago 77 Neighborhoods (100 persons/dot, population data source: 2012-2016 American Community Survey 5-year Estimates).*

The Twitter dataset was used to model the flow pattern. We collected it using the Twitter application programming interface (API). We collected geotagged tweets within Chicago city from Oct.09 to Nov.16, 2016, using the Twitter Stream API. We only kept tweets that fall in the study area. The Twitter users of these tweets include tourists, residents, and commuters near the study area. To compare with the traditional residential segregation, we have to extract the residents' users. We adopt the algorithm in Luo et al. (2016) by checking whether the user's historical activity was clustered within the study area during the evening (8:00 pm-8:00 am).

Therefore, we also collect each user's historic tweets by Twitter timeline API. In the end, we got three million geotagged tweets for the demonstration study.

3.4 Methodology

3.4.1 Construct the time-dependent flow network based on Twitter data

The time-dependent flow network is represented by an origin-destination (OD) flow matrix with a time restriction. Each element of the matrix represents the number of flows between two areal units at time t. We first sequenced each Twitter user's tweets by time. Then, we take each of the two temporal consecutive tweets as one move record, provided that the record satisfies 1) the time interval between two tweets is less than 4 hours and 2) the straight-line distance between two tweets is longer than 100m (Q. Liu, Wang, and Ye 2018). These two restrictions ensure each move record's integrity as much as possible and eliminate the uncertainty of the GPS signal.

After extracting all qualified moves, we use the "points in polygon" operation in QGIS to determine each move's origin and destination neighborhood. Finally, we create a time-dependent OD matrix by aggregating moves according to their origin, destination, and time at the neighborhood level. In the case study, we divide one day into six time slots (4 hours each). In each time slot *t*, we use only the moves that overlap with this time slot to create the OD flow matrix, i.e., OD_t, where the element in row *i* and column *j* at time (slot) *t*.

To provide a global overview of the interactions among the three racial/ethnic groups in the case study, we classified the 77 neighborhoods into four types according to their demographic composition: non-Hispanic Black-majority neighborhood, non-Hispanic White-majority neighborhood, Hispanic-majority neighborhood, and Mixed-neighborhood. A

neighborhood is classified as racially majority if one racial/ethnic group is larger than the other two and accounts for more than 50% of its population; if there are no racially majority groups, it is classified as Mixed-neighborhood. For simplicity of description, only the initials are used below to represent the majority type: B-neighborhood, W-neighborhood, H-neighborhood, and M-neighborhood.

3.4.2 Segregation indices with flow patterns

For consistency, we used the notation from Reardon et al. (2004), but we use the traditional polygon as the basic units (i.e., neighborhoods). Let R denotes the entire study area. Assuming *R* is split into n non-overlapping areal units, which are indexed by *i* or *j*. Suppose there are *M* racial/ethnic groups in *R* (*M* = 3 in the case study), and we index them by *x* or *y*, e.g., Black or White group. Let τ denote population size, and a super-positioned tilde (~) is used to indicate the mixing procedure between the target areal unit and other interacted units. Also, we use the subscript *t* to express the time context. We have:

$\phi_t(i,j)$: flow size from unit *i* to unit *j* at time *t*.

$\omega_t(i,j)$: the standardized interaction weight from unit *i* to unit *j* at time *t*.

$\tau_{i,x,t}$: population size of group *x* in unit *i* at time *t*.

$\tau_{i,t}$: population size in unit *i* at time *t* (note that $\tau_{i,t} = \sum_{x=1}^{M} \tau_{i,x,t}$).

$\tilde{\tau}_{i,x,t}$: population size of group *x* in unit *i* at time *t* after a mixing procedure.

$\tilde{\tau}_{i,t}$: population size in unit *i* at time *t* after a mixing procedure.

$T_{x,t}$: population size of group *x* in *R* at time *ss* (note that $T_{x,t} = \sum_{i=1}^{n} \tau_{i,x,t}$).

The measurement of spatiotemporal segregation requires defining the spatial interactions between all pairs of units in region *R* at time *t*. We assume that the strength of the spatial interactions at time *t* is proportional to $\phi_t(i,j)$. If we set the value of $\phi_t(i,i)$ to 0 for now, then the composite population of unit *i* at time *t* can be expressed as $\tilde{\tau}_{i,t}$.

$$\tilde{\tau}_{i,t} = \frac{1}{\Phi_t(i,\cdot)} \sum_{j=1}^{n} \tau_{i,t}\, \phi_t(i,j) \tag{3.1}$$

Where $\Phi_t(i,\cdot) = \sum_{j=1}^{n} \phi_t(i,j)$ and $\phi_t(i,i) = 0$. By moving $\Phi_t(i,\cdot)$ into the summation notation in Equation (3.1), we obtain:

$$\tilde{\tau}_{i,t} = \sum_{j=1}^{n} \tau_{i,t}\, \omega_t(i,j) \tag{3.2}$$

Where $\omega_t(i,j) = \frac{\phi_t(i,j)}{\Phi_t(i,\cdot)}$, and $\sum_j \omega_t(i,j) = 1$ and $\omega_t(i,i) = 0$. Higher value of $\omega_t(i,j)$ means stronger interactions from unit i to unit j at time t. Equation (3.2) is analogous to the spatial lagged variable, where the target unit's attribute of a positive cluster is positively correlated with its neighbors. Here, the $\omega_t(i,j)$ is equivalent to the elements in the row-standardized spatial weight matrix at time t. However, it is worth noting that in the flow-based spatial weight matrix, the number of elements with values greater than zero (i.e., $\omega_t(i,j)$>0) is much larger than those in the matrix constructed by the contiguity-based method.

Although we set $\omega_t(i,i)$ to zero, the flow-based matrices we have constructed so far are sufficient to compute the spatial dissimilarity index. The computation process does not involve $\omega_t(i,i)$, and readers can see the description of the spatial dissimilarity equation in (Cortes et al. 2020) and (Wong 1993). However, the determination of $\omega_t(i,i)$ becomes crucial for Exposure

and Isolation index. In the literature, the equal weights strategy (Wong, 1998) or the distance decay function strategy (White, 1983) are common ways to determine the $\omega(i,i)$. These two approaches are, to some extent, a compromise in the absence of real interaction data. The resulting $\omega(i,i)$ do not reflect the true weights and does not describe the temporal variation of weights. In this chapter, we propose a hierarchy-based strategy to estimate $\omega_t(i,i)$ to measure the weights more accurately. This strategy has two steps:

Step 1) determining hotspot-level (hierarchy-level), $h_t(i)$, for each unit at time t by the Lorenz curve method (Bassolas,2019). The Lorentz curve method arranges the outflow size of all areal units of all times in ascending order and plots them by normalizing the cumulative number of units (x-axis) vs. the fraction of total flow (y-axis). Then it takes the derivative of the Lorenz curve at (1, 1) and extrapolates it to the point at which it intersects the x-axis to get the threshold. We assign $h_t(i)$ to k-level for the areal units on the x-axis greater than this threshold (k starts from 1, which is the highest hotspot level). Then we remove the units of k-level and recalculate the threshold to obtain *(k+1)*-level. Repeat this process to assign $h_t(i)$ for all areal units for each time until the threshold is close to zero or the hotspot level reaches an upper limit *m*, and the rest of the unassigned units will go to the last level. In the case study, we set the upper limit, *m*, to 6.

Step 2) deriving $\omega_t(i,i)$ for unit i at time t by its hotspot level $h_t(i)$. $\omega_t(i,i)$ is obtained by a piecewise function in which each hotspot level $h_t(i)$ is mapped to an interval of $[\alpha,\beta]$ ($0 \leq \alpha \leq \beta \leq 1$). The $\omega_t(i,i)$ is a value between zero to one representing the level of importance of the target unit population in the mixing procedure. When $\omega_t(i,i)$ moves toward zero, the target unit *i* at a time *t* becomes less influential than its interacted units. When $\omega_t(i,i)$ moves toward one, the target unit population will gradually dominate until the interacting neighbors do not

affect the final population composition. To get the mapping result of $\omega_t(i,i)$, we are using a piecewise function, Equation (3.3)：

$$\omega_t(i,i) = \alpha + (h_{t,i} - 1)\frac{\beta - \alpha}{m - 1} \tag{3.3}$$

Where m is the total hotspot levels. Note that, after getting the new $\omega_t(i,i)$, we need to adjust (shrink) other elements in the same row with $\omega_t(i,i)$ to ensure $\sum_j \omega_t(i,j) = 1$.

After obtaining the flow-based spatial weight matrix at time t, we can then compute the flow-based segregation via Equation (3.4), referred to as Flow-based Spatial Exposure Index (FSxPy). For comparison, we also compute 1) Boundary-based Spatial Exposure Index (BSxPy) (Wong 1993; Cortes et al. 2020) and 2) Distance Decay-based Spatial Exposure Index (DDSxPy) (Morgan 1983). BSxPy and DDSxPy are calculated by Equation (3.5), and they construct the interaction weights by using the length of shared boundaries and the distance decay function, respectively. In addition to the different methods of constructing spatial weight matrices, they do not have temporal subscripts, meaning that these spatial weight matrices do not change with time.

$${}_xP_y(t) = \sum_{i=1}^{n} \frac{\tau_{i,x,t}}{T_{x,t}} \cdot \frac{\tilde{\tau}_{i,y,t}}{\tilde{\tau}_{i,t}} \tag{3.4}$$

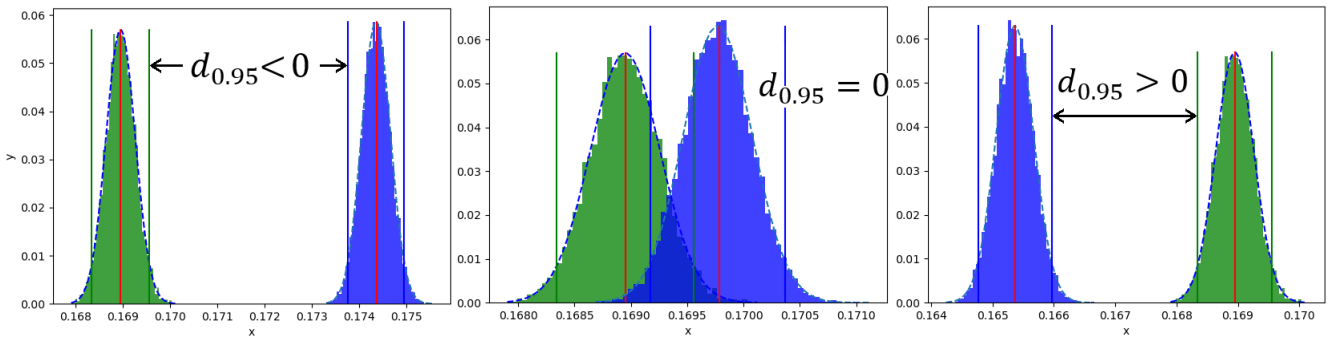
$${}_xP_y = \sum_{i=1}^{n} \frac{\tau_{i,x}}{T_x} \cdot \frac{\tilde{\tau}_{i,y}}{\tilde{\tau}_i} \tag{3.5}$$

Equations (3.4) and (3.5) are global indices, which provide one summarized index for the entire study area. To investigate the local variation of differences, we choose the local variant of spatial exposure index, as shown in Equation (3.6) (Wong and Shaw 2011; Feitosa et al. 2007):

$$ {}_xP_y(i,t) = \frac{\tau_{i,x,t}}{T_{x,t}} \cdot \frac{\tilde{\tau}_{i,y,t}}{\tilde{\tau}_{i,t}} \tag{3.6} $$

3.4.3 Comparative inference

It is a fundamental task to compare the results of different segregation indices mentioned in the previous section. For any set of comparisons, we use the simulation approach to test whether there is a statistical difference. This simulation approach is an extension of the "systematic" inference approach described in (Cortes et al. 2020). The "systematic" approach adds randomness to the current population distribution and calculates segregation for each simulation. In the step of adding randomness, it draws samples from a multinomial distribution, with the success probabilities being the proportions of the population of our interested racial/ethnic group in each areal unit, and with the number of independent trials being the total population of our interested racial group in the entire study area. For any of the indices in the comparison, we conduct 9999 "systematic" simulations, and eventually, we generate two simulating distributions (each contains 9999 simulated results) for one comparison process. Based on the two simulating distributions in each comparison, we fit two Gaussian curves using the maximum likelihood method and calculate confidence intervals under a specified confidence level (see the 95% confidence intervals for the two simulating distributions shown in Figure 3.2). In the end, we define the indices distance between the two simulating distributions as the distance between the confidence intervals, see $d_{0.95}$ in Figure 3.2. If two confidence intervals intersect, then the indices distance is 0, and we say there is no statistical difference between the two indices.



(a) $d_{0.95} < 0$ (b) $d_{0.95} = 0$ (c) $d_{0.95} > 0$

Figure 3.2 *Three Scenarios of Two Index Simulating Distributions (Green Simulating Distribution Denotes the Referenced Index, Blue Simulating Distribution Denotes the Index Used for Comparison). (a) The Indices Distance, $d_{0.95}$, is Less than 0, which Means the Result of the Reference Index is Statistically Significantly Smaller than Its Comparing Index; (b) The Indices Distance, $d_{0.95}$, is Equal 0, which Means There is no Statistically Difference between the Two Comparison Indices; (c) The Indices Distance, $d_{0.95}$, is Greater than 0, which Means the Result of the Reference Index is Statistically Significantly Greater than its Comparison Index.*

3.5 Results

3.5.1 Descriptive analysis of Twitter flow data

Based on the neighborhood classification method in Section 3.4.1, we classified the 77 neighborhoods into four types of neighborhoods (i.e., B-, W-, H- and M-neighborhood). Table 3.1 shows the number, percentage of the total population, and outflow for each type. We see that a substantially larger share of total Twitter outflows originated in the W- than in the B- and H-neighborhoods, with 28% of the population contributing 63% of Chicago's total Twitter outflows. It could be a concrete manifestation of the digital divide in different types of neighborhoods.

Table 3.1 *Basic Statistics of Four Types of Neighborhoods Regarding Count, Percentage of the Population, and Percentage of Twitter Outflows.*

	B-	W-	H-	M-	Total
Neighborhood Counts	28	18	16	15	77
Population Size (%)	694,770 (26%)	770,809 (28%)	557,613 (21%)	691,420 (25%)	2,714,612 (100%)
Twitter Outflow Size (%)	17,262 (7%)	156,085 (63%)	20,069 (8%)	54,499 (22%)	247,915 (100%)

Table 3.2 shows the row-standardized flow percentage between the four neighborhoods' types, the row representing the origin and the column representing the destination. We found that outflows mostly happen between the same type of neighborhoods (diagonal elements are all more than 50%) except for the H-neighborhood. The outflow from H-neighborhood is mainly drawn to the W- and M-neighborhoods. Second, the percentage of outflow is asymmetric across different types of neighborhoods. For example, 14% of total outflow from B-neighborhood goes to W-neighborhoods, but conversely, only 2% of total outflow from W-neighborhood is destined for B-neighborhoods. Similarly, there is a significant asymmetry in the proportion of outflows between H- and W-neighborhoods (36% vs. 4%). W-neighborhoods are the most popular destinations in Chicago city. The asymmetry pattern is consistent with Q. Wang et al.'s finding (2018).

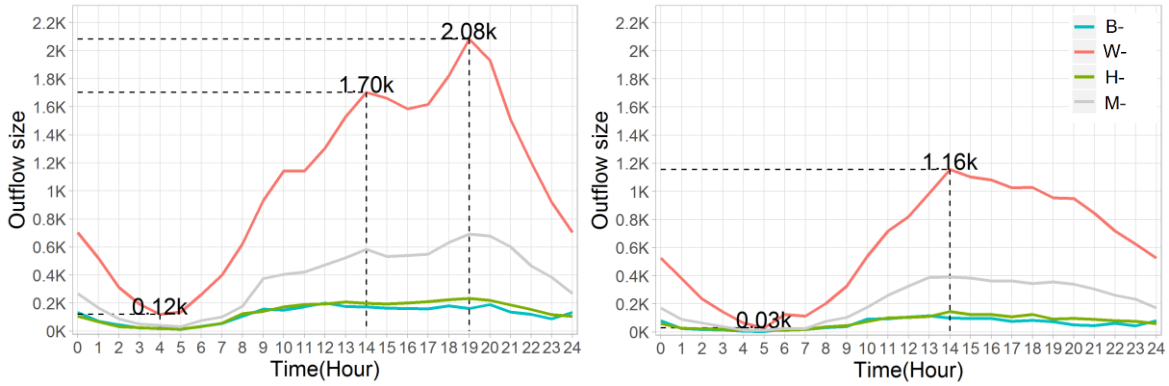
Table 3.2 *The Row Standardized Percentage of flow Size between Four Types of Neighborhoods*

To / From	B-	W-	H-	M-	Total outflow size
B-	59%	14%	13%	14%	17,262 (100%)
W-	2%	83%	4%	11%	156,085 (100%)
H-	10%	36%	29%	25%	20,069 (100%)
M-	4%	33%	9%	54%	54,499 (100%)

Figure 3.3 shows the average outflow size of four neighborhood types on a typical weekday and weekend. Figure 3.4 shows their average travel distance. We find that more flows came from the W-neighborhoods than from the other three types of neighborhoods at any time during a typical weekday or weekend (Figure 3.3a, Figure 3.3b). In contrast, the distance traveled from the W-neighborhoods are significantly shorter than that of the other three types of neighborhoods (Figure 3.4a, Figure 3.4b). This observation is similar to Huang & Wong's finding (2016), which claimed Twitter users of poor communities in Washington D.C. had larger activity spaces than the other wealthier groups. The longer travel distances of the disadvantaged groups are mainly caused by the spatial mismatch of their work location and residence location (Easley 2018). Since the general activity space represented by Twitter data includes work locations, we observed the effects of long work commutes for residents in the B-neighborhoods.

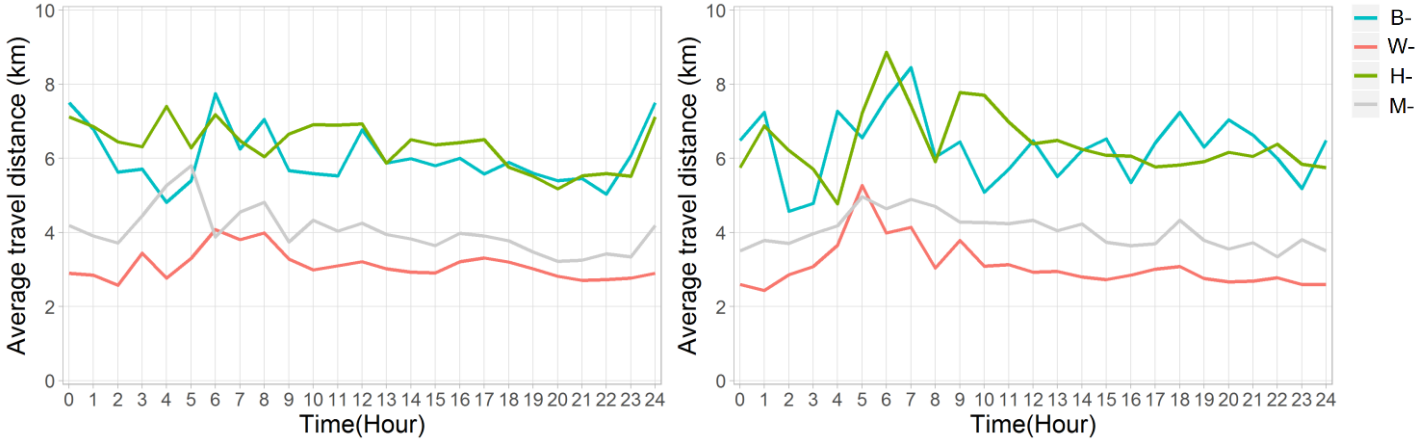
Although the average outflow size and travel distance on weekdays have similar trends with that on weekends, we also see some differences. For example, weekday outflow curves show two flow peaks (at 14:00 and 19:00 in Figure 3.3a) while weekends show only one (at 14:00 in Figure 3.3b). Besides, the weekend maximum peak flow is 44% (i.e., (2.08k-

1.16k)/2.08k) less than the weekday peak flow. This discrepancy of weekday and weekend travel patterns is also reflected in Taxi and CDR datasets (Calabrese and Lorenzo 2011; Zhu et al. 2017). For demonstration purposes in the Chicago case study, we only use the weekday flow data to estimate the interactions between neighborhoods when calculating the FSxPy.



(a) Average Outflow Size on a Typical Weekday (b) Average Outflow Size on a Typical Weekend

Figure 3.3 *Average Outflow Size on Weekday and Weekend by Origin Neighborhood's Type*



(a) Average Travel Distance on a Typical Weekday (b) Average Travel Distance on a Typical Weekend

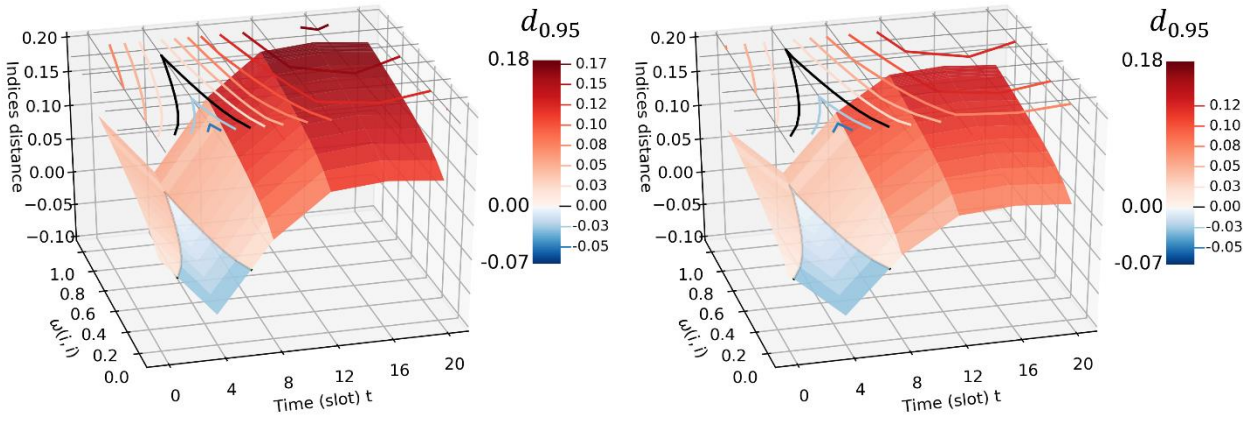
Figure 3.4 *Average Travel Distance in Weekday and Weekend by Origin Neighborhood's Type*

3.5.2 Comparison result of global exposure index

According to Section 3.4.2, we know that the major difference between the global index of FSxPy, BSxPy, DDSxPy is their ways of constructing the spatial weight matrix. Therefore, we can focus on the parameters of construction methods when comparing FSxPy to BSxPy or FSxPy to DDSxPy. We first compare FSxPy to BSxPy of the Black to White group (Figure 3.5). In Figure 3.5a-3.5f, the x- and y-axis represent the time slot t of FSxPy and the parameter of $\omega(i,i)$ of BSxPy, respectively. The z-axis represents their indices distance ($d_{0.95}$) with FSxPy as the reference index. For illustration purposes, we assume that if one unit has the bottom level in the flow hierarchical structure, it has no interactions with other units. So, we set β equal to 1.0, and only change α in FSxPy. Increasing α from 0.0 to 1.0 is the process of weakening the influence of flow data in the weights. Figures 3.5a to 3.5f show how the gradually increasing α affects the indices distance between FSxPy and BSxPy. We use gradient colors from red to blue to represent the high (positive) to low (negative) values of indices distance. We also use the same color to plot contours of indices distance on top of each 3-D figure. But we use a black line to indicate $d_{0.95}$ equal to 0.0, and it means the FSxPy index has the same segregation results as BSxPy.

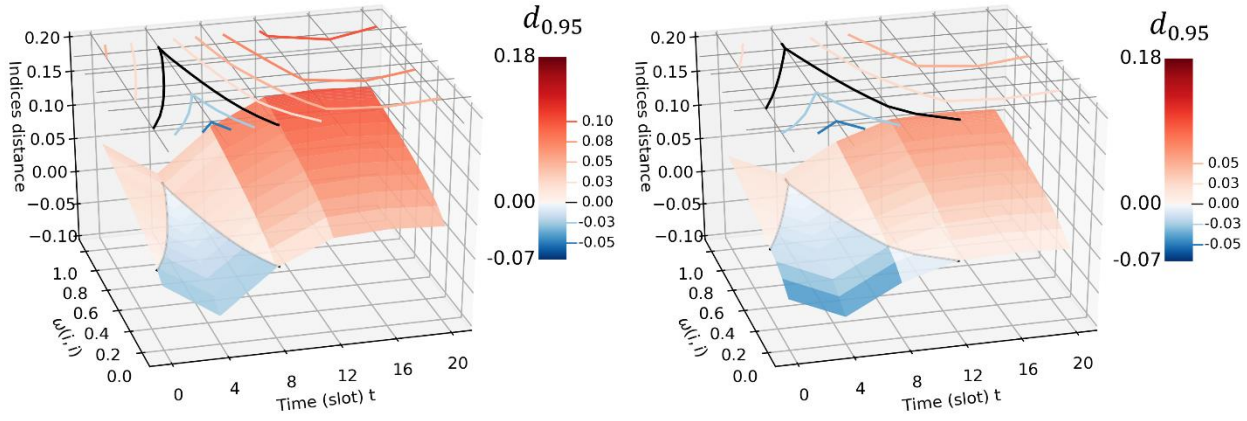
From Figures 3.5a-3.5e, we see a clear temporal pattern along the x-axis (time slots *t*). For any given $\omega(i,i)$ of BSxPy (x-axis), the indices distance, $d_{0.95}$, reaches the lowest value at 4:00 and maximum value at some point between 12:00 and 20:00. This trend is consistent with the outflow pattern in Figure 3.3a. Besides, for any given time slot t (y-axis), a greater indices distance appears when increases $\omega(i,i)$ of BSxPy. From Figure 3.5a to 3.5f contour plots, we find that the area of red contour lines (i.e., $d_{0.95} > 0$) is shrinking until it disappears when $\alpha = \beta = 1.0$ in Figure 3.5f. It shows the process of disappearing flow patterns impacts.

Similarly, we compared FSxPy and DDSxPy of the Black to the White group (Figure 3.6a-3.6f). In DDSxPy, we use the Gaussian kernel with the bandwidth parameter to model the distance decay interactions. We let the y-axis represents the bandwidth in Figure 3.6a-3.6f; the x- and z-axis have the same meanings as in Figure 5. When bandwidth increases, the Gaussian kernel will become flattened, which corresponds to a smaller $\omega(i,i)$ in BSxPy. Thus, the bandwidth of DDSxPy and $\omega(i,i)$ of BSxPy have an inverse effect on $d_{0.95}$. For example, two surfaces in Figure 5a and Figure 3.6a are oriented in different directions. Except for the orientation, the $d_{0.95}$ 3-D surface in Figure 3.6 exhibits the same V-shaped pattern in the time dimension as in Figure 3.5. Besides, we could identify the points where three indices have no differences by following the black contours lines.



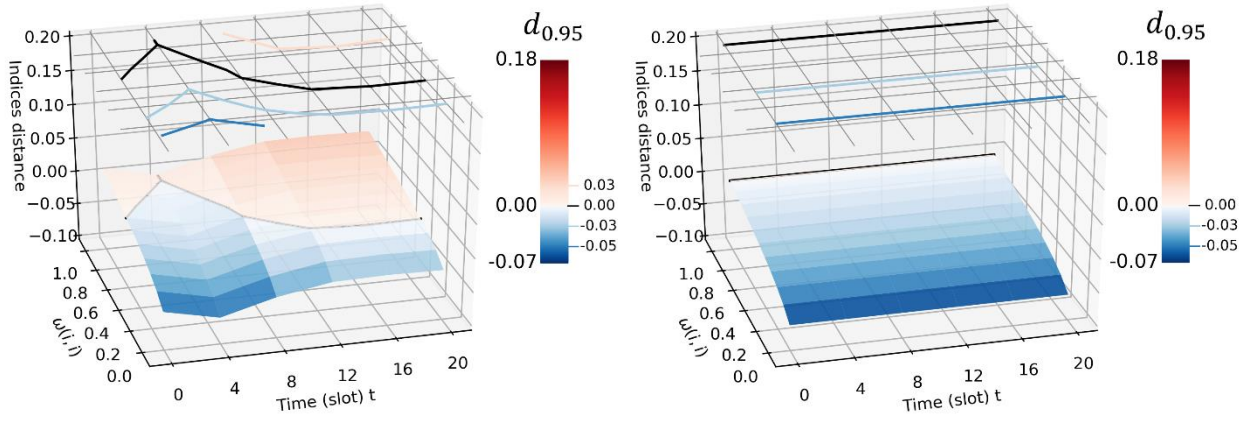
(a) α=0.0 and β=1.0

(b) α=0.2 and β=1.0



(c) α=0.4 and β=1.0

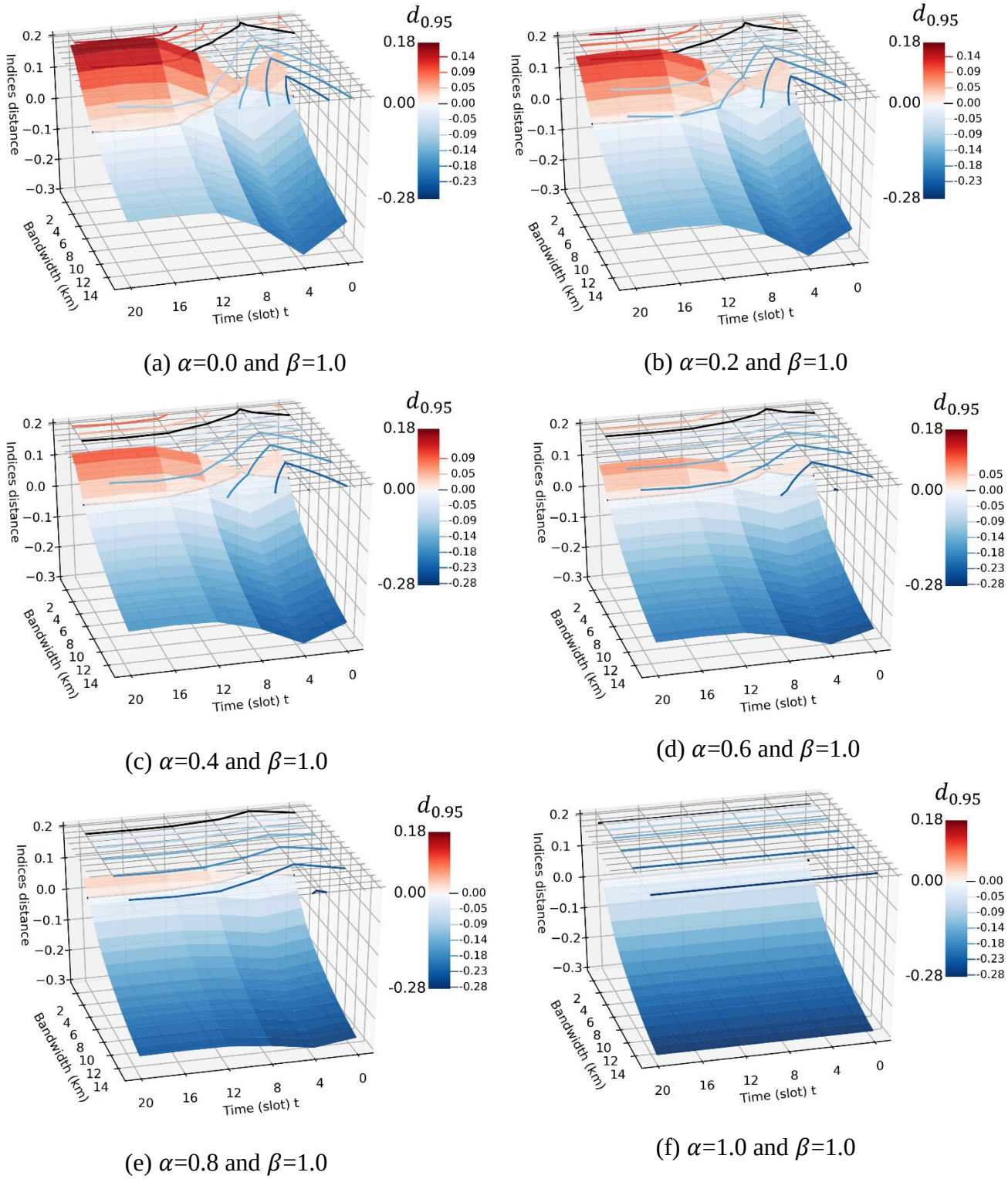
(d) α=0.6 and β=1.0



(e) α=0.8 and β=1.0

(f) α=1.0 and β=1.0

Figure 3.5 *Statistical Difference Between FSxPy (Interested Parameters are α, β,Time (Slot) t) and BSxPy (Interested Parameter is $\omega(i,i)$) of Black to White Group (FSxPy - BSxPy).*



(a) α=0.0 and β=1.0

(b) α=0.2 and β=1.0

(c) α=0.4 and β=1.0

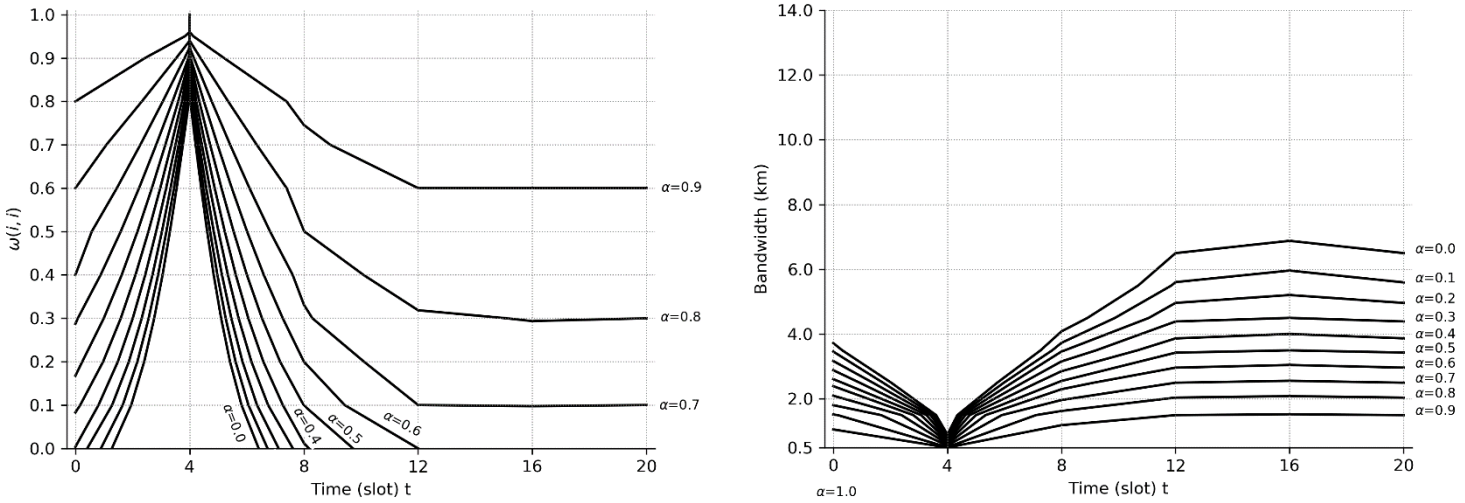
(d) α=0.6 and β=1.0

(e) α=0.8 and β=1.0

(f) α=1.0 and β=1.0

Figure 3.6 *Statistical Difference between FSxPy (Interested Parameters are α, β,Time (Slot) t) and DDSxPy (Interested Parameter is Bandwidth) of Black to White Group (FSxPy - DDSxPy).*

To better examine the effect of the variation of parameter α, we draw the contours of $d_{0.95} = 0$ for each of the six 3-D plots in Figure 3.5 and Figure 3.6 in one plot (see Figure 3.7a and 3.7b). We can see that a singularity is formed at 4:00 when the results of FSxPy are always less than or equal to BSxPy and DDSxPy, regardless of what α to be chosen. It corresponds to the flow data of Figure 3.3a, in which the flow reaches its minimum at 4:00. Population returns to its place of residence at that time, the exchange between areal units reaches its minimum. Accordingly, residential segregation will dominate when the interaction weights of FSxPy reaches a minimum value. However, the weight construction method of using boundary length (BSxPy) or the distance decay function (DDSxPy) ignores these decreasing interactions at 4:00, resulting in a biased (higher positive bias) exposure result.



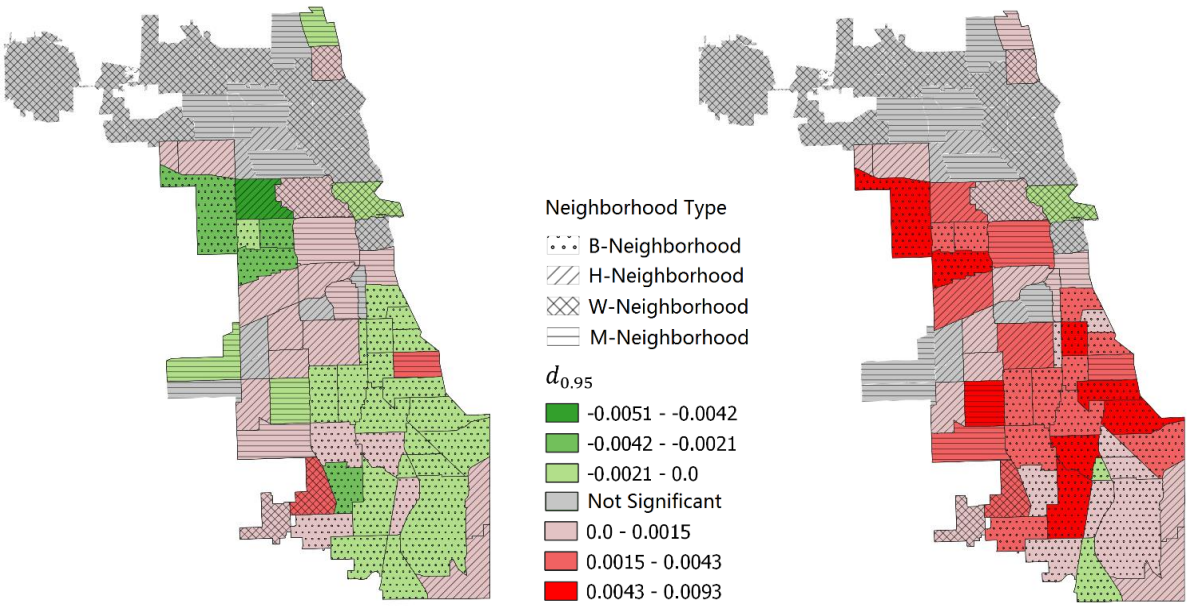
(a) FSxPy vs BSxPy (b) FSxPy vs. DDSxPy

Figure 3.7 *The Contours of* $d_{0.95} = 0$ *Under Different* α *in Different Time with Fixing* $\beta = 1.0$

3.5.3 Comparison result of local exposure index

We employ Equation (3.6) to compare the local version of FSxPy and BSxPy for the Black to White group. We create two maps to show the indices distance heterogeneity at the local scale at the time (slots) 4:00 (2:00-6:00) and 12:00 (10:00-2:00) (Figure 3.8). The colors of green to red represent the difference between FSxPy and BSxPy from low to high, and different infill styles describe the neighborhood types.

In Figure 3.8, we found that the spatial clustering of the negative indices distance regions (green clusters in Figure 3.8a) or positive indices distance regions (red clusters in Figure 3.8b) are almost identical to the distribution of B-neighborhoods. It means that if we use FSxPy to assess segregation of Black to White, the Black group in B-neighborhoods is less exposed to the White group than the results of BSxPy at 4:00 (Figure 3.8a) and is more exposed than BSxPy at 12:00 (Figure 3.8b). In contrast, the indices distance between FSxPy and BSxPy of Black to White was not statistically significant for almost all W-neighborhoods, which means that the flow pattern has little impact on the segregation calculation for W-neighborhoods. The little impact contradicts the fact that the outflow from W-neighborhoods is significantly larger than that of B-neighborhoods (Figure 3.3a) if we assume that more flow leads to more impact on segregation computation. However, considering that users from B-neighborhoods are more likely to travel to W-neighborhoods and the reverse does not stand (Table 3.2), it is not surprising that adding flow patterns have more effect on B-neighborhoods than W-neighborhoods.



(a) 4:00 (b) 12:00

Figure 3.8 *Map of Differences between two exposure index (FSxPy - BSxPy) of Black to White at the time (a) 4:00 and (b) 12:00*

3.6 Discussion

It is reasonable and promising to introduce the population mobility network, hierarchical structure, and temporal information into calculating the conventional segregation indices. The proposed population flow-based method has unique advantages over interaction estimation methods based on topology, geometry properties, or distance decay functions. It can better reflect the impacts of population movements on the degree of segregation in both the temporal and spatial dimensions. First, the V-shaped curve along the time dimension demonstrates the dynamic nature of the temporal dimension in the case study. It is consistent with our expectation of population movement over time. Second, the spatial clustering patterns in Figure 3.8 clearly show the heterogeneity of impacts of population flow in the spatial dimension, and this local

variation is closely related to the type of the neighborhood and its flow patterns. In addition, the choice of parameters significantly influenced segregation results. For example, with a fixed β, in the process of increasing α from 0.0 to 1.0, we observe a progressively shallower V-shaped curve along the time dimension. And by controlling the bandwidth in the distance decay method or the $\omega(i,i)$ in the topology-based method would boost or weaken the difference between FSxPy and the other two conventional segregation indices.

In the case study, we also found that the segregation index (the exposure of Black to the White population) of B-neighborhoods changes significantly over time. But we need to interpret this change behavior with caution. We note that at 12:00, the FSxPy of B-neighborhood is significantly elevated on average, which clearly implies that the exposure of the Black to White is improved. If we conclude that the B-neighborhoods are becoming less segregated at 12:00, we may overlook the other side of the inequality. By combining the flow patterns shown in Section 3.5.1, we can see that reduction in segregation comes at the cost of distance and time for B-neighborhoods users, who have to travel a longer distance to get to other neighborhoods (as shown in Figure 3.4). Therefore, we cannot rely on one dimension of the index alone but need to integrate other dimensions to study the underlying segregation or inequality mechanism.

This chapter demonstrates how to include the population mobility patterns into segregation indices and how the flow-based segregation indices differ from conventional indices. We also conducted a systematic comparison study and provide a baseline for the future segregation study with flow data. But there are still some shortcomings. First, we use a discrete (pairwise) approach to estimate the interaction weights of $\omega_t(i,i)$. The discretization obscures part of the information about the outflow, e.g., outflow size belonging to the same hierarchy

level has the same $h_t(i)$ despite the difference in outflow size. Second, we must note that the spatial scale (level) is also an essential factor, but it is not addressed in our study. If the study units change from a neighborhood level to a county level, each unit's population base increases. And the impact of population flow on the demographic composition of each unit decreases. Therefore, a reasonable α at the county level should be larger than α at the neighborhood level. Finally, we cannot deny the limitations of Twitter data. Biases in the data (e.g., the underrepresentation of older populations) and the failure to associate users with socioeconomic attributes (e.g., we do not know the Twitter user's racial/ethnic information) significantly limit its effectiveness and applications. Also, the digital divide is evident in this study, in which W-neighborhoods have more flow than other types of neighborhoods. Although we do not use absolute numbers of flows but proportions to measure interactions that prevent the impact of the digital divide, the impact on the evaluation of the hierarchical structure is inevitable. In a future study, we want to evaluate its impact by combining it with other flow datasets. Because the method illustrated in this chapter can be easily extended to other datasets, such as Smart Card Data, taxi data, CDR. This approach provides a general framework to support a timely, dynamic response to segregation changes.

3.7 Conclusion

In summary, this chapter proposed a method to add population mobility patterns to the conventional segregation indices to portray the dynamics of segregation over time and space. From the systematic comparison of indices, we demonstrate that neighborhood flow networks, hierarchical structure information, time information, and the choice of parameters are all crucial factors to the segregation estimation. The segregation index calculation is a useful reference to identify segregated populations and areas, but understanding its mechanism needs a multi-

dimensional approach. This study facilitates the broader use of flow data in segregation research and provides a powerful tool for urban planners to gain a more comprehensive understanding of the dimension of segregation dynamics.

CHAPTER 4: REPRESENTATIVE BIAS IN SPATIAL MOVEMENTS AND INTERACTIONS AMONG GEOTAGGED SOCIAL MEDIA FLOWS USING SPATIAL PARTIAL LEAST SQUARE REGRESSION[3]

4.1 Introduction

A massive amount of ambient VGI has inspired much research on human movements (Comito, Falcone, and Talia 2016; Hawelka et al. 2014), population distribution dynamics (Tsou et al. 2018; Deville et al. 2014), and spatio-social networks (Y. Liu et al. 2014; Yin et al. 2017). However, "bigger data are not always better data" given its uncertainty and skewness (Boyd and Crawford 2012). Besides, potential biases may exist in the whole analytical cycle from data collection to analysis. These biases may cause substantial uncertainty that could undermine research findings (Liao et al. 2018; M. Kwan 2016). To eliminate uncertainties and to generalize findings, researchers have increased their focus on evaluating VGI's validity, accuracy, representativeness, and uncertainty in terms of social and behavioral characteristics of users (Li, Goodchild, and Xu 2013).

Commonly observed potential biases include positional accuracy, uneven penetration rate across regions, and unproportioned usage rates across different groups (Liao et al. 2018). Being aware of these biases, how representatives are the tweets that infer social activities and human

[3] This chapter is submitted to the journal of Geographic Analysis for peer review.

mobility is of primary concern to geographers and social scientists (Hargittai and Litt 2011; Steiger et al. 2015). Because studying the representativeness not only helps researchers to generalize their findings but also helps to identify social issues, such as the digital divide (Baginski, Sui, and Malecki 2014) that left behind disadvantaged groups (Z. Wang et al. 2019; Shelton et al. 2014). Here, we classify the unproportioned representative biases into first-order and second-order classes. First-order bias emphasizes the fact that the distribution of data does not always correspond to the distribution of certain types of the geographical phenomenon or the population or groups behind it. Second-order bias highlights that in the representation of movements and spatial interactions. We focus on the second-order bias in this chapter.

There are two methodological challenges in studying second-order bias, including high collinearity among demographic/socioeconomic factors and how to handle the spatial autocorrelation structure in flows. They pose high demands for effective statistical models for such biases. The multicollinearity issue can be addressed using variable selection methods, such as stepwise Ordinary Least Square (OLS) and ridge regression models. Besides, dimension reduction methods, such as Principal Components Regression (PCR) and Partial Least Square Regression (PLSR), can also be used. But variable selection methods may omit important variables and ultimately lead to poor interpretation of the model. Dimension reduction methods preserve all variables' impacts by transforming the original variables to a new variable space. But they are only used in first-order bias studies (L. Li, Goodchild, and Xu 2013) and have not been used for studying second-order representative biases. Spatial autocorrelation in flow data can be adjusted using spatial autoregressive interaction models (LeSage and Pace 2008). However, multicollinearity of the demographic/socioeconomic factors is often accompanied by spatial autocorrelation of flows, further increasing the complexity in modeling the second-order

bias. It raises the need for finding a new approach to combining existing models to address both multicollinearity and spatial autocorrelation simultaneously when assessing second-order biases.

This chapter designs an approach to quantify associations of representative biases of VGI flows and local demographic/socioeconomic factors. Specifically, we design a Spatial Partial Least Square Regression (SPLSR) approach to tackle the spatial autocorrelation of flow data and multicollinearity simultaneously when regressing VGI flow data with highly correlated demographic/socioeconomic characteristics. In remaining of this chapter, we first review relevant literature on the progress of the studies of first-order and second-order representativeness in Section 2. Then we introduce the data and variables used in the case study in Section 3. Afterward, we explain our SPLSR approach in detail in Section 4, with study results presented in Section 5. Finally, we discuss the results and draw conclusions in Sections 6 and 7.

4.2 Literature Review

Representative biases in VGI vary across social groups and regions and are concerned by many studies (Shelton, Poorthuis, and Zook 2015; Liao et al. 2018). Representative biases of VGI are mainly caused by positional accuracy, uneven penetration rates across regions, and the unproportioned usage rates across different socioeconomic and demographic groups (Liao et al., 2018). Among them, unproportioned usage rates are of most concern for researchers and may compromise the robustness of spatial analysis based on VGI. To address this concern, researchers mainly focus on whether the first-order distribution and the second-order interaction obtained from VGI are consistent with the distribution of their corresponding background context.

4.2.1 First-Order Representative Biases

People's digital traces are closely related to their physical activities, so the user-generated content is often connected to individual-based or place-based characteristics and people's engaging activities. In turn, longstanding individual and societal-level inequalities and contextual constraints distort the representativeness of user-generated content (Shelton et al., 2014). For example, well-educated and high-paid young people tend to contribute to more geotagged tweets and photos than other groups in California (Li et al., 2013) or even in the entire US (Wojcik and Hughes 2019).

Spatial heterogeneity, on the other hand, is a manifestation of the uneven spatial distribution of different factors affecting the representativeness of VGI. For instance, Mislove et al. (2011) found a trend of significant underrepresentation of Twitter users in the mid-west and overrepresentation of those in populous counties, after analyzing gender and race/ethnicity distributions among Twitter users across the US. Hecht and Stephens (2014) confirmed the systemic distribution biases of urban and rural areas using Twitter, Flickr, and Foursquare datasets. Their study showed urban areas had five times more geotagged tweets per capita than those from rural areas. Baginski et al. (2014) showed uneven spatial distributions of restaurant reviews between the downtown area of Columbus and surrounding disadvantaged neighborhoods, possibly due to their uneven internet access. However, it should be noted that Baginski's urban-rural digital divide interpretation lacked an analysis of other factors' effects, such as prices and luxury levels of the restaurants and the composition of customer bases.

In short, for first-order biases, both users' characteristics (e.g., wealth or poor) and location context (e.g., urban or rural) affect the ultimate expression of the population they

represent in social media. Consequently, it is critical to consider demographic/socioeconomic factors from different aspects and spatial scales in the VGI study in order to understand its nature, use it appropriately, and generalize the research findings.

4.2.2 Second-Order Representative Biases

Besides the first-order biases, people's travels and interactions also showed skewed distributions over space and time, as found in González et al. (2008) and Jurdak et al. (2015). People spend different amounts of time in different places, and a few favorite destinations are more frequently visited than other places (Song et al. 2010). For example, tourist attractions or recreational areas where residents do not visit routinely have more check-in data. In contrast, business clusters where residents work or visit much more frequently, such as manufacturing centers, are less likely to be popular destinations in check-in data (Y. Hu et al. 2015; García-Palomares, Gutiérrez, and Mínguez 2015). Thus, investigating these potential second-order biases from different perspectives, such as the racial groups, class attributes, and location characteristics, is a prerequisite for promoting a broader application of VGI mobility data.

Representative biases of flow VGI have drawn much attention in transportation studies. It is often treated as hidden variables in predicting traffic flows that used VGI flow data. For example, Lee et al. (2016) predicted the travel demands with an origin-destination (OD) matrix using geotagged tweets. They estimated the inter-regional traffic flows using a Tobit regression model and a latent classes regression model. The representative biases are implicitly included in the latent classes rather than in the model directly. Other studies have tried to verify the extent to which the flow data extracted from VGI can represent the flows from other data sources. Lenormand et al. (2014) performed a comprehensive cross-check analysis to compare

distribution, temporal evolution, and the pattern of flows among Twitter, census, and cell phone data. They found that geotagged Twitter data has lower representativeness than that of mobile phone or census data. However, these three data sources can be interchangeably used at a one-kilometer grids level due to their strong correlations. Phillips et al. (2019) also supported this high correlation conclusion. New data is promising to compensate for shortcomings of traditional survey data (expensive, not up-to-date, limited coverage, etc.). Understanding the relationship between representativeness biases and local characteristics is necessary to better predict traffic demands.

Besides transportation studies, knowledge about second-order biases can also help to directly understand social issues, such as racial segregation and income disparity. Studies showed that socio-spatial segregation was not only caused by a lack of a safe and healthy living environment, access to higher-paying jobs or education, it also ascribed to inequality of social interactions and friendship networks (Echenique and Fryer 2007; Sampson and Levy 2020). Therefore, flow-based networks are gradually becoming a proxy for the social interaction of entire groups between two regions (Phillips et al., 2019). VGI provides a novel way to construct spatially embedded networks and to uncover spatial interactions within different population groups (Shi et al., 2015). However, it remains unclear to what extent the constructed social networks can represent social interaction in a city. To assess the spatial extent and degree of the interaction and isolation, we need further investigations to understand and identify the magnitude of underlying representative biases.

So far, studies on second-order biases have mainly focused on examining correlations in flow volumes and interaction magnitudes across various data sources. There is little research on

which specific users and locations tend to be under- and over-estimated in VGI flows. To model the effects on second-order biases by influencing factors, we need to introduce demographic and socioeconomic variables into the model. However, multicollinearity could emerge when explanatory variables are highly correlated with each other. Though imperfect multicollinearity does not violate Gauss-Markov assumptions, it nevertheless increases the uncertainty of regression coefficients, which can be substantially changed when adding or dropping variables. For example, Naess (2000) attributed the low effects of some characteristics on travel behavior to multicollinearity problems when modeling travel behaviors and location characteristics. Besides multicollinearity, dependences among OD flows become a nonnegligible geographical process. For instance, Chun (2012) compared multiple sets of spatial interaction models and demonstrated that autocorrelation did affect the predictions of interregional commodity flows in the US. Wang et al. (2018) examined friendship connections by using VGI among China's big cities. They found that social connections show spatial dependency in city-level data. Therefore, assuming spatial independence among OD flows or social interactions may be problematic. Recent literature on spatial interaction models has proven the effectiveness of incorporating flow autocorrelation in the model (LeSage and Pace 2008; Chun and Griffith 2011; J. H. Lee, Gao, and Goulias 2015). When studying second-order biases, we would want to explore the association between representative biases in flow data and the different aspects of people and location characteristics. But since multicollinearity and spatial flow dependency exist simultaneously, it may be challenging to obtain a valid model estimation. Therefore, this chapter aims to design an approach that draws on current solutions to the above two methodological problems and considers them simultaneously.

4.3 Study Area and Data

The study area is the 77 Chicago neighborhoods (Chicago Data Portal n.d.), as shown in red polygons in Figure 4.1. Three categories of data are used, including the geotagged tweets, Chicago Household Travel Survey (CHTS), and 20 variables of demographic and socioeconomic characteristics.

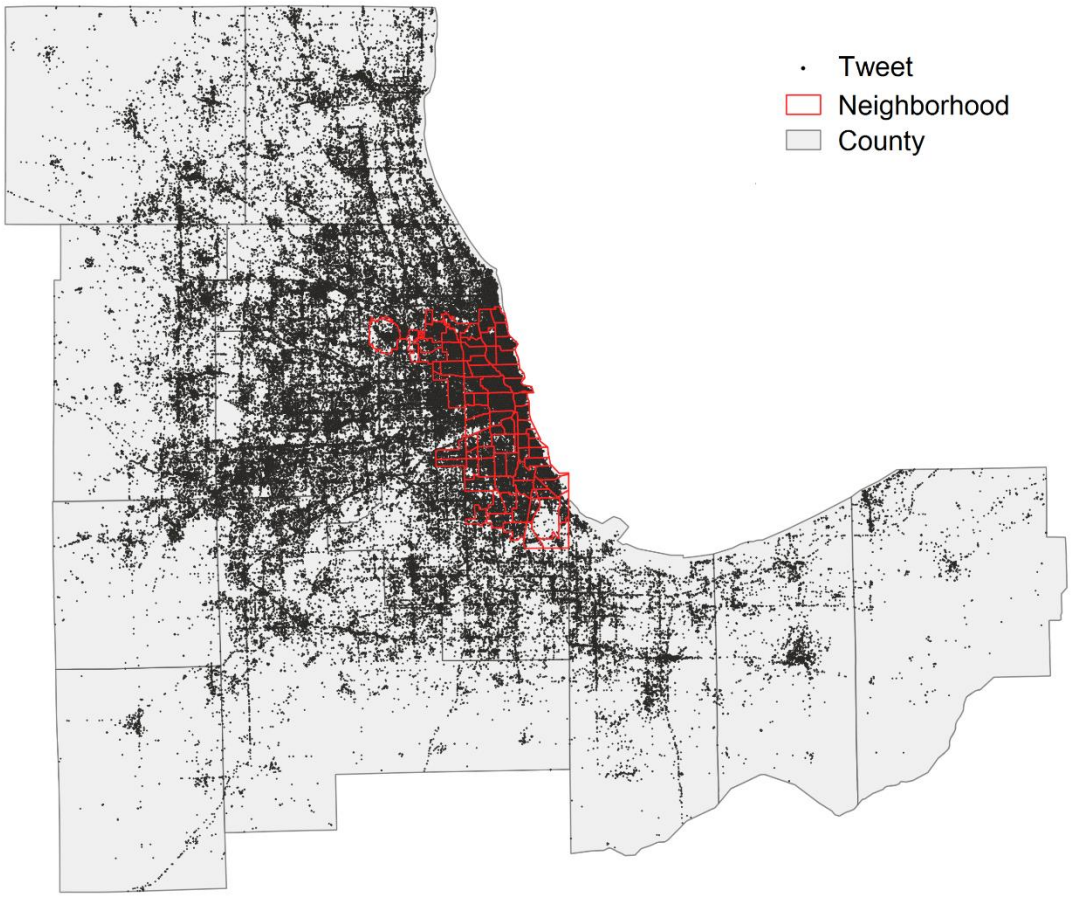


Figure 4.1 *Study Area and Tweets Data Distribution*

Geotagged Tweets

We first collected geotagged tweets in the Chicago metropolitan area from Oct. 9 to Nov. 16, 2016, using the Twitter Stream API. The Twitter users of these tweets included tourists,

residents, and commuters. Their recent historical tweets were collected by Twitter timeline API. But Twitter users were considered valid only if their majority of recent historical geotagged tweets during the evening time (8:00 pm-8:00 am) were located in the ten counties shown in Figure 4.1, which is also the data collecting region of CHTS data. In the case study, we only keep tweets posted by these valid users. In the end, three million geotagged tweets were obtained for the case study (black dots in Figure 4.1).

Our Twitter flow data was extracted from those three million geotagged tweets. We first sequenced tweets by their creation timestamps and then used temporally consecutive tweets to construct each move record. Only move records satisfying the three following conditions were considered in our study: 1) The time interval between two tweets was less than 4 hours; 2) the straight-line distance between two tweets was longer than 100 m (Q. Liu, Wang, and Ye 2018); 3) the move happened in the 77 neighborhoods. The first two conditions ensured each move record's integrity as much as possible and eliminated the GPS signal's uncertainty. By counting the total number of move records between two areal units, we could estimate their flow sizes. In the end, we created a 77×77 square origin-destination (OD) matrix based on the geotagged tweets in which each element represents the flow size between two areal units.

In order to consider digital divide factors in social media data when describing its representativeness, restaurant review counts were also used in our model. We first extracted all reviews for each of 6,044 restaurants in the 77 neighborhoods using "restaurants" as the keyword in Yelp (https://www.yelp.com/). Then we calculated the total review counts for each neighborhood and treated it as a control variable in our regression models.

Chicago Household Travel Survey (CHTS) data

Chicago Household Travel Survey (CHTS) included travel records for a total of 14,390 household members across the ten counties (i.e., gray polygons in Figure 4.1) in the Chicago metropolitan area from Jan 2007 to Mar. 2008. It was created by Chicago Metropolitan Agency for Planning (CMAP) (https://www.cmap.illinois.gov/). CMAP designed an iterative algorithm to weight the raw CHTS surveyed households (14,390) to total households in the ten counties (3,218,100). In this chapter, the weighted flows were used to compare with the geotagged tweets flows. Like processing tweets flow data, we only used the movements within the 77 neighborhoods and created a 77×77 OD flow matrix using CHTS.

Demographic and Socioeconomic Characteristics

We used 20 demographic and socioeconomic characteristics from the 2016 American Community Survey 5-year estimates. They were classified into five categories: race/ethnicity, age, education, occupation, and income. Their details are further described in Table 4.1.

Table 4.1 *Variables Used to Regress the Representative Biases in Twitter Flow Data*

Category	Variable Name	Description
Race/ethnicity	R_HISPANIC	Hispanic population in each areal unit
	R_BLACK	Non-Hispanic African American population in each areal unit
	R_WHITE	Non-Hispanic White population in each areal unit
Age	AGE0_20	Population aged 0-20 years in each areal unit
	AGE20_40	Population aged 20-40 years in each areal unit
	AGE40_60	Population aged 40-60 years in each areal unit
	AGE60_80	Population aged 60-80 years in each areal unit
Education	EDU_LT9TH	Population 18 years over and less than 9th grade in each areal unit
	EDU_HIGH	Population 18 years over and high school graduate in each areal unit
	EDU_A&B	Population 18 years over and Associate's or Bachelor's degree in each areal unit
	EDU_G&P	Population 18 years over and Graduate or professional degree in each areal unit
Occupation	OCC_MBSA	Population of Workers 16 years and over in management, business, science, and arts occupations
	OCC_SER	Population of Workers 16 years and over in service occupations
	OCC_SO	Population of Workers 16 years and over in sales and office occupations
	OCC_NCM	Population of Workers 16 years and over in natural resources, construction, and maintenance occupations
	OCC_PTMM	Population of Workers 16 years and over in production, transportation, and material moving occupations
Income	IN_LT2	Households whose income in the past 12 months (in 2016 inflation-adjusted dollars) were less than $20,000
	IN2_6	Households whose income in the past 12 months (in 2016 inflation-adjusted dollars) were between $20,000-$60,000
	IN6_10	Households whose income in the past 12 months (in 2016 inflation-adjusted dollars) were between $60,000-$100,000
	IN10_15	Households whose income in the past 12 months (in 2016 inflation-adjusted dollars) were between $100,000-$150,000

4.4 Methodology

4.4.1 Spatial Dependence Structure for Flows

In the study area of n areal units, we first used the queen contiguity method to create the typical row-standardized n × n areal spatial weight matrix W. Then, we created a flow spatial dependency structure based on origins and destinations and their neighborhoods. LeSage and Pace (2008) showed three types of flow dependency structures, including origin-based spatial dependence W_o (Equation (4.1)), destination-based spatial dependence W_d (Equation (4.2)), and origin-to-destination spatial dependence W_w (Equation (4.3)).

$$W_o = W \otimes I_n \tag{4.1}$$

$$W_d = I_n \otimes W \tag{4.2}$$

$$W_w = W_o \cdot W_d \tag{4.3}$$

where $\otimes$ denotes Kronecker product, $\cdot$ denotes dot product, and I_n is an $n \times n$ identity matrix. Among three classes of flow dependence structures, W_o captures structures that the origin and its neighbors are likely to create similar flows to its destinations. It reflects the intuition that forces leading to flows from any origin to a particular destination region may create similar flows from neighbors of this origin to the same destination (LeSage and Pace 2008). W_d captures structures that the destination and its neighbors are liked to attract similar flows from its origins. It reflects the intuition that forces leading to flows from an origin to a destination may create similar flows to nearby or neighboring destinations (LeSage and Pace 2008). The similar attractiveness of the destination's neighbors could enhance or diminish flows to this destination. Based on W_o and W_d, W_w reflects the similar flows from neighbors of the origin to neighbors of the destination. Some recent literature also discussed the formation and usage of these three

spatial flow structures (Z. Wang et al. 2018; Chun, Kim, and Kim 2012; J. H. Lee, Gao, and Goulias 2015). In practice, since W_w contains the most comprehensive information about the flow structure, including both the origin and destination units' dependency structure, we choose W_w to represent the flow spatial dependency structure in the case study.

4.4.2 Modeling Representative Bias in Twitter Flow

Step One: Regress Tweets-Based OD Flow on Survey-Based OD Flow.

Given an $n \times n$ OD flow matrix, we can produce an $n^2 \times 1$ vector by stacking its n columns. In this way, we can create y_{Tweet} and y_{CHTS} from the flow matrices created based on geo-tagged tweets and CHTS flow, respectively (here $n = 77$). In the case study, y_{Tweet} and y_{CHTS} used in the regression model (see Equation (4.4)) are log-transformed to make them close to a normal distribution (Chun, Kim, and Kim 2012; Z. Wang et al. 2018). We can regress y_{Tweet} on y_{CHTS} to identify the proportion of Twitter flow data that can be explained by CHTS flow data.

$$y_{Tweet} = \beta_0 + \beta_1 y_{CHTS} + \varepsilon_{ols} \tag{4.4}$$

Using an OLS regression, residuals ε_{ols} represent the part of y_{Tweet} that cannot be explained by y_{CHTS}. It contains the systematic representative bias that we are concerned with since we assume CHTS is the representativeness of true population flows. When the residual of unit *i* to unit *j* is positive, there is a potential over-representativeness of Twitter flow data from *i* to *j*; conversely, if it is negative, it indicates an under-representativeness flow from *i* to *j*. In the next step, we explore the relationship between each demographic and socioeconomic variable and this unexplained portion and identify which variable contributed to over-/under-representativeness.

Step Two: A Spatial PLSR Model

In our analysis, step two is to find the relationship between the residuals in Equation (4.4) and each demographic and socioeconomic variable discussed in Section 3. Typically, the coefficients (effects) of explanatory variables in *X* on *y* can be modeled by multiple linear regression, see Equation (4.5):

$$y = \hat{\varepsilon}_{ols} = X\beta + \xi, \xi \sim N(0, I\sigma^2) \quad (4.5)$$

where y is the $\hat{\varepsilon}_{ols}$ in step one, X is a design matrix for explanatory variables. Current spatial interaction model specifications treat intra-flows and inter-flows separately (Chun, Kim, and Kim 2012). We also split the design matrix X into X_{inter} and X_{intra} and model them separately. X_{inter} is an $n^2 \times 2p$ matrix, its first *p* columns denote explanatory variables of origin units, and last *p* columns denote explanatory variables of the destination units. X_{intra} is an $n \times p$ matrix, include *p* explanatory variables in all *n* units. And dependent variable $\hat{\varepsilon}_{ols}$ can be represented by an $n^2 \times 1$ vector for the inter-flow model and an $n \times 1$ vector for the intra-flow model.

As we discussed in Section 4.2, highly correlated variables and spatial dependency in flows can cause methodological difficulties in the model. PLSR is used to overcome the multicollinearity problem. It aims to reduce both exploratory variables and the response variable to a smaller number of uncorrelated principal components that characterize most of their covariance. It is a useful and effective statistical tool for analyzing the relationship between a response variable and a set of strongly collinear exploratory variables. Since PLSR does not consider the spatial autoregressive process in flows, we design a Spatial PLSR (SPLSR) approach, incorporating a spatial autoregression model into PLSR.

First, we can perform a standard PLSR. Z-Score is used to standardize the design matrix, and the response's mean can be used to center y. The PLSR algorithm can be performed through an interactive procedure to get the principal components. Its detail can be referred to (Mevik and Wehrens 2007; Bennett and Embrechts 2003). After standard PLSR, we can get all *M* principal components (PCs) ($M = 2p$ for inter-flow model and $M = p$ for intra-flow model), and then we need to choose *m* PCs (m ≪ M) to approximate the original design matrix as Equation (4.6):

$$X = T_{(m)}V_{(m)}^{\top} + u_X \tag{4.6}$$

where X is X_{inter} for the inter-flow model and X_{intra} for the intra-flow model. $T_{(m)} = [\overrightarrow{PC_1}, \overrightarrow{PC_2}, \dots, \overrightarrow{PC_m}]$ is the first m PCs or scores. $V_{(m)} = [\overrightarrow{v_1}, \overrightarrow{v_2}, \dots, \overrightarrow{v_m}]$ denotes the orthogonal loading vectors. u_X is the combination of the remaining PCs.

Several approaches can be used to determine *m*, such as Cross-Validation (CV), the Akaike Information Criterion (AIC), and the Bayesian Information Criterion (BIC) (Nengsih et al. 2019). Here, we choose CV techniques, and m is selected when Root-Mean-Square Error (RMSE) reaches the minimum (as suggested in James et al., 2007, pp. 233) or when the cumulative percentage of PC variance meets some chosen threshold. For example, if *m* PCs have explained more than 90% of X before reaching the minimum RMSE, these *m* PCs will be selected.

Second, we regress $\hat{\varepsilon}_{ols}$on these selected m PCs. It can be expressed by Equation (4.7)

$$\hat{\varepsilon}_{ols} = T_{(m)}\beta^* + \xi^* \tag{4.7}$$

where $\beta^* = V_{(m)}^{\top}\beta$ and $\xi^* = u_X\beta + \xi,\ \xi^* \sim N\left(0, I{\sigma^*}^2\right)$. Equation (4.7) are estimated by OLS and estimation of β^* can be expressed as Equation (4.8).

$$\widehat{\beta^*} = \left(T_{(m)}^{\top} T_{(m)}\right)^{-1} T_{(m)}^{\top} \hat{\varepsilon}_{ols} \tag{4.8}$$

$$\hat{\beta} = \left(V_{(m)}^{\top}\right)^{-1} \widehat{\beta^*} = V_{(m)} \left(T_{(m)}^{\top} T_{(m)}\right)^{-1} T_{(m)}^{\top} \hat{\varepsilon}_{ols} \tag{4.9}$$

$$Var\left(\hat{\beta}\right) = V_{(m)} Var\left(\widehat{\beta^*}\right) V_{(m)}^{\top} \tag{4.10}$$

In SPLSR, we can add a procedure for checking spatial autocorrelation of residuals ξ^* in Equation (4.7). Moran's I test can be used to do this. Its detailed procedure can be found in Anselin and Bera (1998). If spatial dependence is not detected, then results in Equation (4.9) and (4.10) are our final estimation of β and its variance for Equation (4.5). Otherwise, we would use a Spatial Autoregressive Model (SAR) to regress $\hat{\varepsilon}_{ols}$ on those m PCs, as shown in Equation (4.11). This is because it would involve spatial lags of the dependent variable (Chun, Kim, and Kim 2012; Z. Wang et al. 2018; LeSage and Pace 2008).

$$\hat{\varepsilon}_{ols} = \rho W_w \hat{\varepsilon}_{ols} + \alpha\iota + T_{(m)}\beta^* + \tilde{\xi} \tag{4.11}$$

$$\tilde{\xi} \sim N(0, I\sigma_s^2)$$

W_w is the origin-to-destination spatial dependence matrix (see Equation (4.3)). For intra-flow data W_w is simplified to the usual spatial weight matrix W. ρ is a scalar spatial dependence parameter. ι is a vector of ones, with associated scalar parameter α. Vector $\tilde{\xi}$ denotes a normally distributed disturbances with zero mean and constant variance.

The complicated flow dependence structure increases the difficulty of interpreting the model estimates. In this framework, we can adopt the summary measure of impacts proposed by LeSage and Pace (2009 pp. 39), which is the average total impact ($\overline{M}_{total}$) as Equation (4.12).

$$\bar{M}_{total} = (1 - \hat{\rho})^{-1}\widehat{\beta^*_{SAR}} \quad (4.12)$$

where, $\hat{\rho}\ and\ \widehat{\beta^*_{SAR}}$ are estimations for Equation (4.11).

After getting the average total impact $\bar{M}_{total}$, we can use it to calculate the estimation of $\hat{\beta}$ in Equation (4.5), as shown in Equation (4.13):

$$\hat{\beta} = V_{(m)}\bar{M}_{total} = V_{(m)}(1 - \hat{\rho})^{-1}\widehat{\beta^*_{SAR}} \quad (4.13)$$

In order to draw inferences regarding the statistical significance of $\hat{\beta}$ in Equation (4.13), we would be required to construct its distribution, which can be constructed by a large number of simulations (9,999 simulations were done in the case study). In each simulation, we would draw parameter values from the multivariate normal distribution implied by $\hat{\rho}$ and $\widehat{\beta^*_{SAR}}$ and their variances that can be estimated from Equation (4.11), and then we can use them to compute one $\hat{\beta}$ value. After forming an empirical distribution of $\hat{\beta}$, we can use a t-test to assess its statistical significance.

4.5 Results

4.5.1 Result of Inter-Flow Model

Following the analytical procedures described in the last section, Table 4.2 shows the results in step one of the inter-flow model. The coefficient is statistically significant, which indicates a positive relationship between Twitter flow data and CHTS flow data. However, adjusted-R^2 is not high (0.217), around 80% of the variation in Twitter flows is left unexplained after partial out CHTS flows.

Table 4.2 *OLS Inter-Flow Model in Step One*

	Log(Tweets Inter-flow)	
	Coef.	Std. Error
Constant	1.667	0.128***
Log(CHTS flow)	0.320	0.019***
Adjusted R^2	0.217	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Step two started with the 80% unexplained variations, and it was considered as the new y in the standard PLSR model. Seven out of 42 PCs were selected, and they explained 91.4% of the variation in X_{inter} and 35.7% of the variation in y. After a linear regression (Equation (4.7)) and Moran's I test, spatial autocorrelation was detected in Equation (4.7). Thus, SPLSR was used to model $\hat{\varepsilon}_{ols}$ over the seven PCs. For comparison, Table 4.3 also shows the results of a standard PLSR model and a stepwise OLS model. To observe the degree of multicollinearity in the stepwise OLS model, we calculated VIF values for each variable in the stepwise OLS model.

Table 4.3 *Inter-Flow Models in Step Two*

	SPLSR			Standard PLSR			Stepwise OLS			
Variables	Coef.	Std. Error		Coef.	Std. Error		Coef.	Std. Error		VIF
Origin										
R_HISPANIC	0.288	0.030	***	0.192	0.016	***	0.338	0.105	**	17.1
R_BLACK	-0.213	0.022	***	-0.141	0.012	***				
R_WHITE	-0.098	0.017	***	-0.058	0.010	***	0.351	0.189	.	55.2
AGE0_20	-0.321	0.043	***	-0.177	0.025	***	-0.250	0.107	*	17.9
AGE20_40	0.046	0.016	**	0.057	0.009	***				
AGE40_60	0.049	0.018	**	0.008	0.010					
AGE60_80	-0.069	0.021	**	-0.075	0.011	***	0.457	0.144	**	32.3
EDU_LT9TH	0.149	0.020	***	0.105	0.011	***	0.274	0.113	*	19.9
EDU_HIGH	-0.065	0.015	***	-0.069	0.008	***	-0.212	0.139		30.1
EDU_A&B	0.017	0.007	*	0.023	0.004	***	-2.092	0.526	***	428.2
EDU_G&P	0.005	0.013		0.026	0.008	***	-2.074	0.427	***	282.9
OCC_MBSA	0.084	0.008	***	0.068	0.004	***	3.739	0.655	***	664.8
OCC_SER	0.159	0.022	***	0.086	0.013	***				
OCC_SO	-0.069	0.011	***	-0.039	0.007	***	-0.653	0.291	*	131.6
OCC_NCM	-0.315	0.035	***	-0.203	0.019	***	-0.459	0.106	***	17.5
OCC_PTMM	-0.024	0.007	**	-0.018	0.004	***				

Table 4.3 (continued)

IN_LT2	0.064	0.021	**	0.062	0.012	***				
IN2_6	0.164	0.020	***	0.095	0.011	***				
IN6_10	0.113	0.014	***	0.066	0.008	***	0.788	0.286	**	127.1
IN10_15	-0.015	0.008	.	-0.002	0.005					
#YELP_REVIEW	0.207	0.019	***	0.148	0.010	***				
Destination										
R_HISPANIC	0.220	0.024	***	0.161	0.013	***	0.334	0.101	**	16.0
R_BLACK	-0.199	0.020	***	-0.142	0.011	***				
R_WHITE	-0.073	0.015	***	-0.037	0.009	***	0.336	0.170	*	45.0
AGE0_20	-0.356	0.047	***	-0.195	0.027	***	-0.436	0.085	***	11.2
AGE20_40	0.059	0.019	**	0.072	0.010	***				
AGE40_60	0.022	0.017		-0.008	0.010					
AGE60_80	-0.100	0.028	***	-0.105	0.015	***	0.631	0.148	***	33.9
EDU_LT9TH	0.115	0.018	***	0.085	0.010	***				
EDU_HIGH	-0.052	0.018	**	-0.064	0.010	***	-0.319	0.149	*	34.4
EDU_A&B	0.028	0.009	**	0.036	0.005	***	-2.650	0.531	***	436.7
EDU_G&P	0.012	0.011		0.028	0.006	***	-1.801	0.403	***	251.2
OCC_MBSA	0.098	0.009	***	0.081	0.004	***	4.759	0.777	***	935.6
OCC_SER	0.180	0.022	***	0.105	0.013	***	0.145	0.098		15.0
OCC_SO	-0.022	0.007	**	-0.007	0.004					
OCC_NCM	-0.341	0.038	***	-0.215	0.021	***	-0.401	0.100	***	15.5
OCC_PTMM	-0.018	0.007	*	-0.012	0.004	**	0.328	0.137	*	29.0
IN_LT2	0.081	0.017	***	0.060	0.010	***				
IN2_6	0.268	0.033	***	0.155	0.018	***				
IN6_10	0.075	0.009	***	0.047	0.005	***				
IN10_15	-0.067	0.014	***	-0.029	0.008	**	-0.692	0.234	**	85.2
#YELP_REVIEW	0.123	0.010	***	0.091	0.005	***				

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

From Table 4.3, we noticed a great difference between estimates of stepwise OLS compared to those of the other two models. First, the multicollinearity of stepwise OLS was of concern, as EDU_A&B, EDU_G&P, OCC_MBSA had abnormally large VIF values. We also calculated the Pearson correlation coefficient values and found that the correlation between EDU_A&B and OCC_MBSA reached 0.99, and the correlation between EDU_G&P and OCC_MBSA reached 0.97. Figure 4.2 shows the Pearson correlation coefficient plots between

different variables. From Figure 4.2, we can see that high education level is associated with OCC_MBSA occupation, and the White population (R_WHITE) is also significantly positively associated with high income (IN10_15), high education level (EDU_A&B), and OCC_MBSA occupation. Also, in the stepwise OLS model, signs of R_WHITE, AGE60_80, EDU_A&B, and OCC_PTMM are reversed compared to those of the other two models. It is reasonable to suspect that multicollinearity may lead to inconsistencies in these signs. In contrast, by comparing SPLSR and PLSR, we can see that their estimates maintain sign consistency but have differences in the magnitude of $\hat{\beta}$ estimations. Therefore, we argue that both PLSR and SPLSR work well to avoid the effects of high collinearity among their explanatory variables.

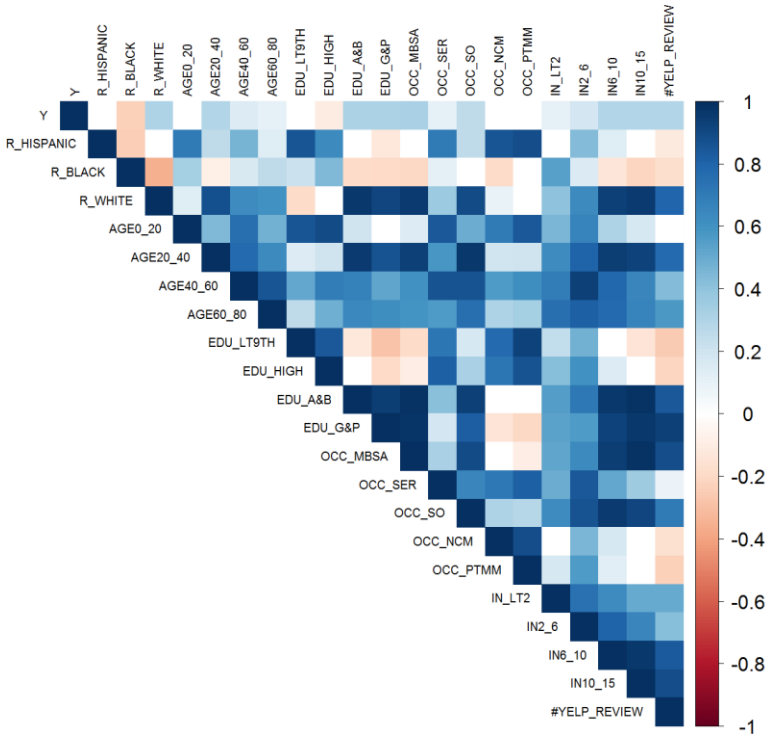


Figure 4.2 *Correlation Coefficients Map*

In the following, we mainly analyzed the impacts of each coefficient in SPLSR on representativeness biases. In the race/ethnicity category, we found a positive effect of the R_HISPANIC variable (0.288*** of origin and 0.220*** of destination), whereas R_BLACK and R_WHITE are negatively impacting the biases. This suggests that the Hispanic population tends to generate excess flows than expected in inter-neighborhoods travel. In contrast, the travels of the White and African American people tend to be underrepresented. In the age category, we found that AGE20_40 (0.046** of origin and 0.059** of destination) and AGE40_60 (0.049** of origin and 0.022 of destination) positively contribute to the representative biases; however, AGE0_20 (-0.321*** of origin and -0.356*** of destination) and AGE60_80 (-0.069** of origin and -0.100*** of destination) have a negatively impact. This result is consistent with our expectations because the very young (0-20) and old (≥60) populations have limited access to Twitter. In first-order biases studies, AGE0_20 and AGE60_80 age groups are also inherently unrepresentative (L. Li, Goodchild, and Xu 2013); thus, their movement can easily be underestimated in Twitter. For the education and occupation category, we found that only OCC_MBSA and OCC_SER showed a significant and high positive effect; EDU_A&B, EDU_G&P have a small impact on biases even though they have high correlations with OCC_MBSA and OCC_SER. We conjecture that impacts of EDU_A&B, EDU_G&P have been captured by those two occupational variables above. Surprisingly, the contribution of low and middle income (IN2_6 and IN6_10) to representativeness bias is strongest, rather than that of the high-income variable. The #YELP_REVIEW variable is a control variable, and its positive effect is in line with our expectations.

Because our design matrix was z-score normalized, the sum of absolute values of significant coefficients in the same category can reflect the magnitude of impacts of each category. By calculating the sum of absolute coefficients of each category, we found impact

sequence of categories to be: occupation category (1.311) > racial/ethnic category (1.092) > age category (1.001) > income (0.847) > education category (0.426). Clearly, occupational factors dominate the representative bias of inter-neighborhoods tweets flows in Chicago. Although the education category comes last, it does not mean it has weak prediction ability. It only implies that it does not provide additional information for predicting bias after including its positively correlated occupation variables.

4.5.2 Result of Intra-Flow Model

The intra-flow model's step one results are shown in Table 4.4. They are consistent with the inter-model's step one results in Table 4.2, where CHTS data explains approximately the same amount of information in the Twitter data.

Table 4.4 *OLS Intra-Flow Model in Step One*

	Log(Tweets Intra-flow)	
	Coef.	Std. Error
Constant	1.122	1.041
Log(CHTS flow)	0.480	0.108***
Adjusted R^2	0.211	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

After the standard PLSR, we selected four PCs to approximate the design matrix X_{intra} in Equation (4.7). The four PCs explained 94.58% of explanatory variables and 48.75% of y variable. As there is no evidence of spatial autocorrelation in Equation (4.7) using Moran's I test, a standard PLSR was used in step two for the intra-flow model. We also performed the stepwise OLS for comparison purposes. Their results are displayed in Table 4.5.

Table 4.5 *Intra-Flow Models in Step Two*

	Standard PLSR				Stepwise OLS				
Variables	Coef.	Std. Error	Pr(>|z|)		Coef.	Std. Error	Pr(>|z|)		VIF
R_HISPANIC	-0.148	0.042	0.000	***	-1.351	0.388	0.001	**	21.5
R_BLACK	-0.301	0.104	0.004	**	-0.404	0.180	0.028	*	4.6
R_WHITE	0.075	0.013	0.000	***					
AGE0_20	-0.244	0.078	0.002	**					
AGE20_40	-0.055	0.038	0.156						
AGE40_60	0.128	0.039	0.001	**					
AGE60_80	0.038	0.025	0.126						
EDU_LT9TH	-0.016	0.017	0.338						
EDU_HIGH	-0.024	0.012	0.055	*	-0.489	0.340	0.156		16.5
EDU_A&B	-0.024	0.036	0.513		-3.022	1.149	0.011	*	188.6
EDU_G&P	0.021	0.025	0.395		-1.176	0.707	0.101		71.5
OCC_MBSA	0.031	0.021	0.137		3.554	1.386	0.013	*	274.2
OCC_SER	0.315	0.109	0.004	**	0.583	0.335	0.086	.	16.0
OCC_SO	-0.102	0.053	0.057	.					
OCC_NCM	-0.152	0.046	0.001	**					
OCC_PTMM	0.048	0.036	0.182		0.933	0.480	0.056	.	33.0
IN_LT2	0.236	0.081	0.004	**					
IN2_6	0.107	0.034	0.002	**	0.622	0.391	0.117		21.9
IN6_10	0.076	0.009	0.000	***					
IN10_15	-0.004	0.031	0.892						
#YELP_REVIEW	0.260	0.065	0.000	***	0.650	0.260	0.015	*	9.7

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

The stepwise OLS model in Table 4.5 also shows large VIF values of OCC_MBSA and EDU_A&B, similar to the inter-flow model. In the following, we mainly discussed the results of the standard PLSR model. First, the coefficients of variable R_HISPANIC and R_WHITE were reversed in sign from those in the inter-flow model. This reversion suggests that the Hispanic group had more inter-neighborhoods flows exposed to Twitter than intra-neighborhood flows. Second, like in the inter-flow model, variables AGE0_20 and AGE40_60 still showed a respectively negative and positive impact on representative biases. However, AGE20_40 and AGE60_80 estimations became statistically insignificant. Third, the coefficient of education

variables was greatly reduced in the intra-model, and they became insignificant in almost all of them. However, the income category's impacts did not change much comparing to their impacts in the inter-model. Overall, using the sum of average absolute significant coefficients method, we found that impact sequence became occupation category (0.569) > race/ethnicity category (0.524) > income category (0.419)> age category (0.372) > education category (0.024). It showed that most categories remain in the same order as the inter-model impact sequence.

4.6 Discussion

We designed an SPLSR approach to examine the relationship between local demographic/socioeconomic factors and second-order representative biases of flows obtained from geotagged social media data. SPLSR model was carried out in two steps. The first step assessed how strong the explanatory power of one flow data was for another flow data. In the Chicago case, the CHTS data explained only about 20% of the Tweets flow in either inter-model or intra-model. This explanatory power of this level was significantly lower than that of tweets in other studies. But it might be caused by the relatively short collection duration of Tweets data in the case study. We plan to use more data to verify the relationship between tweets and CHTS travel flow in the future.

The second step analyzed how each demographic/socioeconomic factor affected the representative biases we found from step one. Based on this case study of Twitter flow data in Chicago, we found that the occupation is the most influential category in both intra-model and inter-model. Besides, variables in the same category showed different impacts on representativeness bias in both intra-model and inter-model. For example, populations in management, business, science, arts, and service occupations positively affected representative

biases. In contrast, populations in natural resources, construction, maintenance, production, transportation, and material moving occupations showed a negative impact on representative biases. Among age category variables, the neighborhoods with a high proportion of population aged 20-60 were likely to experience overrepresented Twitter flows, but a high proportion of aged 0-20 and ≥60 would lead to under-represented. Moreover, the same variables could have similar impacts on the representative bias in both inter- and intra-models. These include R_BLACK, AGE0_20, AGE40_60, EDU_HIGH, OCC_SER, OCC_SO, OCC_NCM, IN_LT2, IN2_6, IN6_10, and #YELP_REVIEW. Among them, groups in African American (R_BLACK), less than 20 years old (AGE0_20), having high-school education (EDU_HIGH), in sales and office occupations (OCC_SO), and in natural resources, construction, and maintenance occupations (OCC_NCM) were at risk of under-representation in both models while other groups were over-represented. Furthermore, the same variables might have even opposite effects on representative biases in intra- and inter-models, such as R_HISPANIC and R_WHITE. It confirmed the necessity to compare inter- and intra-flow models or analyze them separately in the representative bias study. In our case study, the Hispanic group variable showed a positive effect in the inter-model but a negative effect in the intra-model. The White group is the other way around. One possible explanation for flipping signs is that the Hispanics were more likely to share movement between neighborhoods than within neighborhoods, while the White groups reversed to Hispanic. But this speculation needs further verification.

The proposed SPLSR moved beyond the conventional multiple linear regression method because it integrated the spatial autoregressive process into PLSR and could also avoid multicollinearity, even adding more variables. With the case study of Chicago, this chapter verified the necessity and validity of adding the spatial autoregressive process of flow data into

PLSR. In the inter-flow model, Moran's I test detected a significant spatial dependence in residuals. Thus, SPLSR was used to reduce estimation bias in PLSR. In this chapter, we only considered the spatial dependence between flows. However, many factors could generate spatial dependence, such as spatial heterogeneity of explanatory variables (Anselin and Bera 1998) or the omitted variables with spatial dependences in the model (LeSage and Pace 2009, pp.28). The underlying flow spatial dependency mechanism might be an interesting topic to explore in the future.

Mobility data represents not only mobility directions but also connections or interactions between regions and ethnic groups. Understanding where or under what conditions mobility data is more prone to generate representative biases is a prerequisite for better generalizing current human mobility and social interaction findings. The work in this chapter makes a necessary inquiry on representations of mobility data. The designed SPLSR approach links multiple socioeconomic attributes and representational biases, and it investigates segregation and social inequality to some extent. For example, why did the R_HISPANIC variable produce more flow bias than the R_WHITE variable in the inter-flow model? Why was the R_BLACK variable under-represented in both models? Was it because African Americans had less access to social media platforms or just didn’t want to share? Understanding the flow biases of geotagged data would provide insight into these questions. This chapter starts approaching these social issues, but more studies may be needed to answer these questions better.

There are several limitations in the designed SPLSR method and analysis carried out in the case study. First, our framework did not deal with the problem of a large number of zero-valued flows. When many zero-valued flows occurred, the model specification had to be

changed accordingly. For example, Poisson estimation can be used for spatial autoregressive models (LeSage, Fischer, and Scherngell 2007; Lambert, Brown, and Florax 2010). Second, CHTS flow data were assumed to be underlying true inter- and intra-neighborhoods flows when modeling the representative biases of Twitter flow data. However, we did not know to which extent it could reflect real flows. Therefore, the representative bias in Twitter flow data could also be caused by bias in CHTS survey data. We suggest using multiple sources of data to produce a more comprehensive analysis of representative bias.

4.7 Conclusion

When using data extracted from VGI, we should always be aware of its existing representative biases. In this study, the SPLSR approach was designed to simultaneously consider the spatial autocorrelation of flows and the high correlation of socioeconomic factors in representative bias analysis. It was also used to analyze the relationship between the second-order representative biases of the Twitter flow data and local demographic and socioeconomic characteristics in Chicago. The case study results verified that user groups with different characteristics in occupation, age, race, income, and education have different influences on their participation in social media platforms and had different impacts on the representative biases in Twitter data. However, more research could be needed to explain from a psychological and sociological perspective why people with some specific social and demographic attributes were more likely to show their mobility in social media platforms. And additional flow data from other study area or other sources have to be included to verify whether our representative bias findings in Chicago applies to other regions.

CHAPTER 5: CONCLUSIONS AND FUTURE WORK

5.1. Conclusions

With the increasing convenience of transportation and communication technologies, there is a great spatial mismatch between people's residential spaces and their daily out-of-home activity spaces (Horner and Mefford 2007). The analysis of people's activity spaces and mobility patterns is essential for inferring the potential interactions between places of different population groups. These inferred interactions have drawn increasing attention in socio-spatial segregation studies (Wong 2002; Farber et al. 2015).

This dissertation aims to understand and quantify socio-spatial segregation from a dynamic perspective based on people's activity spaces and mobility patterns. This study has four interconnected components, including the analytical framework's design, comparing mobility patterns between different groups, the integration of VGI flow data in conventional segregation indices, and the representative evaluation of VGI flow data.

Chapter 2 reconstructed and compared the activity spaces and mobility patterns of different groups of visitors and residents. I found that social media data can help us to get a more detailed user group classification. Based on Twitter users' origin place, I classified visitors into state, national and international groups. Then, I constructed the county-level mobility network for each user group and investigated the centrality of these networks. The results showed that short-distance travels are the main activity type for all groups in the study area. Moreover, the

centrality index of mobility networks for each user group showed a core–peripheral structure, and it is positively related to the total population size of each county. To mitigate MAUP, I split the study areas into hexagon grids at different spatial scales, including 1km-grid, 2.5km-grid, 5km-grid, and 10km-grid. I analyzed the entropy evenness index of each group of users in each hexagon grid. The spatial distribution of entropy evenness index at each spatial scale confirmed the core–peripheral mobility networks structure.

In Chapter 3, I studied the methodology to improve the exposure segregation index using Twitter data. The designed approach effectively incorporates the population flow network, hierarchical structure, and temporal information obtained from Twitter flows. In the demonstration case study, I computed the global segregation of black to white groups. After extensive simulations and comparisons, we confirmed the designed global exposure segregation index could better reflect the impacts of population movements on the degree of segregation in both the temporal and spatial dimensions. Especially, it can reflect temporal changes of the population movement over the course of a day. In contrast, the conventional indices cannot reflect the spatial and temporal dynamic changes because they rely only on geometric and adjacency information.

I also calculated the designed exposure index at the neighborhood level and their differences with the conventional segregation index. These differences showed clear spatial clustering patterns. The clustering patterns indicated the heterogeneity of impacts of population flow in the Chicago city. At 4:00 am, most African American neighborhoods had a lower exposure index than the conventional one. The lower index indicates conventional fixed spatial interaction strategies overestimate the interactions that exist in human mobility data. In contrast,

at 12:00 o'clock, African American neighborhoods experienced an overall increase in exposure index, which means they had more chance of meeting peoples of white-dominated neighborhoods. However, they experienced another dimension of inequality at this time. We found that residents of African American neighborhoods travel a much greater distance than those of whites neighborhoods did in Twitter flow data. Thus, it reminded us that we should portray them from different perspectives to get more accurate assessments of disparity and socio-spatial segregation.

Besides, I compared the clustering of local segregation changes with the distribution of neighborhood types. My analysis showed that the spatial clustering of the negative regions (at the time of 4:00) or positive regions (at the time of 12:00) were almost identical to the distribution of African Americans-dominated neighborhoods. This consistent distribution suggested that temporal changes of segregation in African Americans dominated neighborhoods are masked by conventional indices of segregation. Therefore, our proposed approach to fusion population mobility patterns into social segregation index has unique advantages over the conventional ones in the spatio-temporal dimension.

Finally, I found that both the conventional indices and our proposed segregation index are affected by their parameter configurations. For example, for my proposed segregation index, with a fixed parameter β, in the process of increasing parameter α from 0.0 to 1.0, I observed a progressively shallower V-shaped curve along the time dimension. The conventional segregation indices are affected by the bandwidth in the distance decay method or the $\omega(i,i)$ in the topology-based method. The result reminds us to be aware of the algorithm uncertainty (M. Kwan 2016).

Nowadays, we have the unprecedented computational power and an improved software interface, and researchers can complete complex calculations by simply setting up input data and parameters with a few mouse clickings. But the algorithms become black boxes, not to mention to investigate its uncertainties. Although researchers benefited from the powerful and convenient computation, they were inadvertently bearing the algorithms' uncertainty and parameter settings. Therefore, I emphasize the need to pay close attention to algorithms and parameters' uncertainty when implementing socio-spatial segregation analysis programs. For example, the software may include an interface that provides researchers with the ability to show changes in results at different parameter settings and customized algorithms.

Chapter 4 mainly analyzed the effects of five categories of demographic and socioeconomic characteristics on second-order representative biases of VGI. I designed an SPLSR approach to overcome the challenges of multicollinearity and spatial flow autocorrelation in representativeness biases study. The SPLSR uses PLSR to handle the multicollinearity problem and uses a spatial autoregressive model to address the spatial autocorrelation problem in the population flow data extracted from VGI. Based on a case study of Twitter flow data in Chicago, I revealed the most influential categories of variables for the flow biases in intra-neighborhood and inter-neighborhood, respectively. Results show that 1) occupations contributed the most to representational biases in inter- and intra-neighborhood flows. 2) Population of African Americans less than 20 years old, having high-school education, and in sales and office, natural resources, construction, and maintenance occupations are at risk of being underrepresented. 3) Same factors can have different or even opposite effects on the representativeness bias in inter- and intra-neighborhood flow model. Hence, it is necessary to compare inter- and intra-flow models or analyze them separately in future studies.

5.2. Limitations in the Research Method

In the discussion section of Chapters 3, 4, and 5, I have separately discussed each study's limitations. I will not repeat them here. This section mainly illustrated the limitations of VGI, algorithms, and models from a more broad perspective. The uncertainties of data, algorithms, and model propagate and accumulates throughout the analysis cycle. It is of great challenge to assess the accumulation and propagation of uncertainties.

First, VGI data has many limitations. For example, it is noisy, incomplete sampling, unorganized and incomplete data, etc. These shortcomings could be addressed or alleviated by collecting more data or using more elaborate data cleaning methods. Alternatively, researchers can work with the corresponding companies to obtain more comprehensive data. Another important barrier to preventing the broader use of VGI for socio-spatial segregation is that they are seldom associated with users' socioeconomic attributes. One could easily determine the home location of social media users and infer some neighborhood socioeconomic data from that, but doing so undoubtedly commits an ecological fallacy and introduces uncertainties to the following analysis. Unfortunately, this is an unavoidable limitation of this study, as well as other studies using VGI data. The question of how we should measure the amount of uncertainty in the data is still an unresolved issue.

Second, algorithms and models may also introduce uncertainties. Such uncertainties are easily ignored by researchers, resulting in unpredictable outcomes. In Chapter 3, I used a large number of simulations to investigate the impacts of algorithm parameters on results. I showed that the results from my newly designed and conventional methods could be numerically equal by artificially adjusting their input parameters. Our study confirmed that their parameter settings

influence both my designed method and conventional methods. However, I have not proposed an effective method to determine which parameters are appropriate for a particular study yet. The parameter setting is closely related to the object and environment of one study. Local knowledge of the study area must guide the choice of parameter settings. Thus, this also limits the analysis to other metropolitan areas.

Besides, the choice of statistical model also brings uncertainties to this study. In Chapter 4, I addressed the autocorrelation of the flow data by a spatial autocorrelation model. However, there are many specifications of spatial autocorrelation models. The three most commonly used models are the spatial lag model, spatial error model, and spatial Durbin model. There is no simple answer as to which model is appropriate for one flow data. Though I adopted the spatial lag model in the demonstration study, this does not mean that we exclude the possibility of other models.

In addition to uncertainties, the case studies are limited by the spatial scale in the analysis. In Chapter 2, I examined the mobility patterns of various groups at the County level due to the limit of data quality. I found that my data becomes too sparse to build interaction networks when analyzing population movement in smaller units, but the county level appears too coarse for city planners. A standard solution nowadays is to increase the time duration to collect data and accumulate more data to avoid sparsity. We need to note that our analysis framework does not limit the scale of analysis. Researchers can choose different scales for their analysis depending on the data availability and their research topics.

Finally, I want to emphasize that more work needs to be done on multi-source data fusion. I have already highlighted the importance of multi-source data fusion in Chapters 3 and

4. However, in the three case studies, the Twitter data is the only new data source due to the limitation of our data availability. The single data source imposes many limitations on our research. First of all, I do not know whether our findings can be generalized to other geotagged big data. For example, VGI can be obtained from many other platforms (e.g., Flickr, Yelp), which are different from the Twitter data used in this dissertation. The mobility data can also be obtained from other sources, such as taxi data, bus card data, and cell phone data. How much difference they will make in the segregation evaluation is yet still unknown. Secondly, I do not know whether our findings can be generalized to other regions. It calls for the need to synthesize multiple open big data to improve the generalization of our study.

5.3. Future Work

Based on the limitations I mentioned in the previous section. I argue that it is vital to validate the framework from multiple data, multi-scale, and multi-location perspectives for the segregation study. These perspectives enable the algorithms and models to be more generalizable. Secondly, I expect to design an evaluation system that can qualitatively and quantitatively evaluate the uncertainties throughout the entire analysis cycle.

The designed analytical framework provides many analytical functions for social segregation study, but this dissertation only focuses on part of them. In the future, I want to study or implement more functions designed in this framework. I also want to improve the framework with the development of new technology, new studies, and new data. For example, I plan to integrate more big data processing or analysis methodologies into this framework to process VGI data more efficiently. I also want to design a big data management system to store and manage

the used data and analytical results so that they can provide reference information for other researchers.

The analysis in this dissertation mainly used the geotagged information of social media users. In the future, I want to use machine learning algorithms to extract more information from the rich text and photos in social media data. While users expose their location, they also expose what they were feeling, seeing, or thinking at that moment. The contents posted by users are the result of their interaction with the local environment at a given time. The text and photo information can also be very valuable for segregation study. But this information is trivial and unstructured, which impedes its usage on a large scale. Even though there are many algorithms for image recognition and machine learning recently, the analysis of unstructured text and photos seems almost impossible without manual intervention. In the future, we need to use or design new machine learning algorithms to automatically classify or recognize the emotion or contextual information in the text and photos.

In this dissertation, I underlined the application of human mobility data in the evaluation of socio-spatial segregation. More specifically, the study extends conventional segregation evaluation methods so that it can accommodate new VGI data. However, beyond segregation evaluation, the methodologies in this dissertation can also be applied to a wider range of fields, such as public health, business, tourism, and other scientific fields. It is important to note that we also need to use knowledge from other disciplines to interpret population movement patterns, processes, and mechanisms. Only with the integration of multidisciplinary knowledge can we correctly and comprehensively explain the dynamics and mechanisms behind group segregation under consideration of the population movements.

REFERENCE

Ahn, Michael J., and Bob McKercher. 2015. "The Effect of Cultural Distance on Tourism: A Study of International Visitors to Hong Kong." *Asia Pacific Journal of Tourism Research* 20 (1): 94–113. https://doi.org/10.1080/10941665.2013.866586.

Amaral, Francisco, Teresa Tiago, and Flávio Tiago. 2114. "User-Generated Content: Tourists' Profiles on TripAdvisor." *International Journal of Strategic Innovative Marketing* 01 (December): 137–47. https://doi.org/10.15556/IJSIM.01.03.002.

Amini, Alexander, Kevin Kung, Chaogui Kang, Stanislav Sobolevsky, and Carlo Ratti. 2014. "The Impact of Social Segregation on Human Mobility in Developing and Industrialized Regions." *EPJ Data Science* 3 (1): 6. https://doi.org/10.1140/epjds31.

Andereck, Kathleen L., Karin M. Valentine, Richard C. Knopf, and Christine A. Vogt. 2005. "Residents' Perceptions of Community Tourism Impacts." *Annals of Tourism Research* 32 (4): 1056–76. https://doi.org/10.1016/j.annals.2005.03.001.

Andriotis, Konstantinos, and Roger D. Vaughan. 2003. "Urban Residents' Attitudes toward Tourism Development: The Case of Crete." *Journal of Travel Research* 42 (2): 172–85. https://doi.org/10.1177/0047287503257488.

Ankomah, Paul K, John L Crompton, and Dwayne Baker. 1996. "Influence of Cognitive Distance in Vacation Choice." *Annals of Tourism Research* 23 (1): 138–50. https://doi.org/10.1016/0160-7383(95)00054-2.

Anselin, Luc, and Anil K Bera. 1998. "Spatial Dependence in Linear Regression Models with an Introduction to Spatial Econometrics." In *Handbook of Applied Spatial Analysis*, 237–89. CRC Press.

Apparicio, Philippe, Joan Carles Martori, Amber L Pearson, Éric Fournier, and Denis Apparicio. 2014. "An Open-Source Software for Calculating Indices of Urban Residential Segregation." *Social Science Computer Review* 32 (1): 117–28.

https://doi.org/10.1177/0894439313504539.

Aslam, Salman. 2017. “Twitter by the Numbers: Stats, Demographics & Fun Facts.” 2017. https://www.omnicoreagency.com/twitter-statistics/.

Baginski, James, Daniel Sui, and Edward J Malecki. 2014. “Exploring the Intraurban Digital Divide Using Online Restaurant Reviews: A Case Study in Franklin County, Ohio.” *The Professional Geographer* 66 (3): 443–55. https://doi.org/10.1080/00330124.2013.866431.

Bassolas, Aleix, Hugo Barbosa-Filho, Brian Dickinson, Xerxes Dotiwalla, Paul Eastham, Riccardo Gallotti, Gourab Ghoshal, et al. 2019. “Hierarchical Organization of Urban Mobility and Its Connection with City Livability.” *Nature Communications* 10 (1): 4817. https://doi.org/10.1038/s41467-019-12809-y.

Batra, Adarsh. 2009. “Senior Pleasure Tourists: Examination of Their Demography, Travel Experience, and Travel Behavior Upon Visiting the Bangkok Metropolis.” *International Journal of Hospitality & Tourism Administration* 10 (3): 197–212. https://doi.org/10.1080/15256480903088105.

Batty, Michael. 2002. “Thinking about Cities as Spatial Events.” *Environment and Planning B: Planning and Design* 29 (1): 1–2. https://doi.org/10.1068/b2901ed.

———. 2010. “Space, Scale, and Scaling in Entropy Maximizing.” *Geographical Analysis* 42 (4): 395–421. https://doi.org/10.1111/j.1538-4632.2010.00800.x.

———. 2013. *The New Science of Cities*. MIT press.

Bauder, Michael, and Tim Freytag. 2015. “Visitor Mobility in the City and the Effects of Travel Preparation.” *Tourism Geographies* 17 (5): 682–700. https://doi.org/10.1080/14616688.2015.1053971.

Bennett, Kp, and Mj Embrechts. 2003. “An Optimization Perspective on Kernel Partial Least Squares Regression.” *Nato Science Series Sub Series III Computer and Systems Sciences* 190: 227–50.

Boyd, Danah, and Kate Crawford. 2012. “Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon.” *Information, Communication &*

Society 15 (5): 662–79. https://doi.org/10.1080/1369118X.2012.678878.

Brands, Jelle, Tim Schwanen, and Irina van Aalst. 2014. "Spatiotemporal Variations in Nightlife Consumption: A Comparison of Students in Two Dutch Cities." *Applied Geography* 54 (October): 96–109. https://doi.org/10.1016/j.apgeog.2014.07.008.

Bromley, Rosemary D.F., Andrew R Tallon, and Colin J Thomas. 2003. "Disaggregating the Space-Time Layers of City-Centre Activities and Their Users." *Environment and Planning A* 35 (10): 1831–51. https://doi.org/10.1068/a35294.

Browning, Christopher R, and Brian Soller. 2014. "Moving Beyond Neighborhood: Activity Spaces and Ecological Networks As Contexts for Youth Development." *Cityscape (Washington, D.C.)* 16 (1): 165–96. http://www.ncbi.nlm.nih.gov/pubmed/25105172%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4121985.

Buliung, Ron N., and Pavlos S. Kanaroglou. 2006. "Urban Form and Household Activity-Travel Behavior." *Growth and Change* 37 (2): 172–99. https://doi.org/10.1111/j.1468-2257.2006.00314.x.

Calabrese, Francesco, and Giusy Di Lorenzo. 2011. "Estimating Origin- Destination Flows Using Mobile Phone Location Data." *Cell* 10: 36–44. https://doi.org/10.1109/MPRV.2011.41.

Candia, Julián, Marta C. González, Pu Wang, Timothy Schoenharl, Greg Madey, and Albert-László Barabási. 2008. "Uncovering Individual and Collective Human Dynamics from Mobile Phone Records." *Journal of Physics A: Mathematical and Theoretical* 41 (22): 224015. https://doi.org/10.1088/1751-8113/41/22/224015.

Carter, S.M., & West M.A. 1998. "Urban Tourism and Visitor Behavior." *Small Group Research* 29 no. 5: 583–601. https://doi.org/0803973233.

Chen, Cynthia, Jason Chen, and James Barry. 2009. "Diurnal Pattern of Transit Ridership: A Case Study of the New York City Subway System." *Journal of Transport Geography* 17 (3): 176–86. https://doi.org/10.1016/j.jtrangeo.2008.09.002.

Chen, Yimin, Xiaoping Liu, Xia Li, Xingjian Liu, Yao Yao, Guohua Hu, Xiaocong Xu, and

Fengsong Pei. 2017. “Delineating Urban Functional Areas with Building-Level Social Media Data: A Dynamic Time Warping (DTW) Distance Based k-Medoids Method.” *Landscape and Urban Planning* 160 (December): 48–60. https://doi.org/10.1016/j.landurbplan.2016.12.001.

Cheng, Z, S Jian, M Maghrebi, TH Rashidi, and ST Waller. 2020. “Integrating Household Travel Survey with Social Media Data to Improve the Quality of OD Matrix: A Comparative Study” 21 (6): 2628–36. https://trid.trb.org/view/1572462.

Chhabra, Deepak. 2007. “Ethnicity and Marginality Effects on Casino Gambling Behavior.” *Journal of Vacation Marketing* 13 (3): 221–38. https://doi.org/10.1177/1356766707077691.

Chicago Data Portal. n.d. “Boundaries-Neighborhoods.” Accessed December 4, 2020. https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Neighborhoods/bbvz-uum9.

Chua, Alvin, Loris Servillo, Ernesto Marcheggiani, and Andrew Vande Moere. 2016. “Mapping Cilento: Using Geotagged Social Media Data to Characterize Tourist Flows in Southern Italy.” *Tourism Management* 57 (December): 295–310. https://doi.org/10.1016/j.tourman.2016.06.013.

Chun, Yongwan, and Daniel A. Griffith. 2011. “Modeling Network Autocorrelation in Space–Time Migration Flow Data: An Eigenvector Spatial Filtering Approach.” *Annals of the Association of American Geographers* 101 (3): 523–36. https://doi.org/10.1080/00045608.2011.561070.

Chun, Yongwan, Hyun Kim, and Changjoo Kim. 2012. “Modeling Interregional Commodity Flows with Incorporating Network Autocorrelation in Spatial Interaction Models: An Application of the US Interstate Commodity Flows.” *Computers, Environment and Urban Systems* 36 (6): 583–91. https://doi.org/10.1016/j.compenvurbsys.2012.04.002.

Comito, Carmela, Deborah Falcone, and Domenico Talia. 2016. “Mining Human Mobility Patterns from Social Geo-Tagged Data.” *Pervasive and Mobile Computing* 33 (December): 91–107. https://doi.org/10.1016/j.pmcj.2016.06.005.

Cortes, Renan Xavier, Sergio Rey, Elijah Knaap, and Levi John Wolf. 2020. *An Open-Source*

Framework for Non-Spatial and Spatial Segregation Measures: The PySAL Segregation Module. *Journal of Computational Social Science*. Vol. 3. Springer Singapore. https://doi.org/10.1007/s42001-019-00059-3.

Crowder, Kyle, and Liam Downey. 2010. "Interneighborhood Migration, Race, and Environmental Hazards: Modeling Microlevel Processes of Environmental Inequality." *American Journal of Sociology* 115 (4): 1110–49. https://doi.org/10.1086/649576.

Crucitti, Paolo, Vito Latora, and Sergio Porta. 2006. "Centrality Measures in Spatial Networks of Urban Streets." *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 73 (3): 036125. https://doi.org/10.1103/PhysRevE.73.036125.

Deville, Pierre, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R. Stevens, Andrea E. Gaughan, Vincent D. Blondel, and Andrew J. Tatem. 2014. "Dynamic Population Mapping Using Mobile Phone Data." *Proceedings of the National Academy of Sciences* 111 (45): 15888–93. https://doi.org/10.1073/pnas.1408439111.

Dharmowijoyo, Dimas B E, Yusak O Susilo, and Anders Karlström. 2014. "Day-to-Day Interpersonal and Intrapersonal Variability of Individuals' Activity Spaces in a Developing Country." *Environment and Planning B: Planning and Design* 41 (6): 1063–76. https://doi.org/10.1068/b130067p.

Duncan, Otis Dudley, and Beverly Duncan. 1955. "A Methodological Analysis of Segregation Indexes." *American Sociological Review* 20 (2): 210. https://doi.org/10.2307/2088328.

E. Eric Boschmann. 2008. "Getting to Work: A Mixed Methods Analysis of Metropolitan Area Working Poor Employment Access." The Ohio State University.

Eagles, Paul F.J., Peter a. Johnson, Luke R. Potwarka, and Chelsea Parent. 2015. "Travel Distance Classes for Tourism Destinations: A Proposal from Ontario Provincial Park Camping." *Journal of Ecotourism* 14 (1): 64–84. https://doi.org/10.1080/14724049.2015.1071829.

Easley, Janeria. 2018. "Spatial Mismatch beyond Black and White: Levels and Determinants of Job Access among Asian and Hispanic Subpopulations." *Urban Studies* 55 (8): 1800–1820. https://doi.org/10.1177/0042098017696254.

Echenique, Federico, and Roland G Fryer. 2007. “A Measure of Segregation Based on Social Interactions.” *The Quarterly Journal of Economics* 122 (2): 441–85. https://doi.org/10.1162/qjec.122.2.441.

Fang Bao, Ya, and Bob Mckercher. 2008. “The Effect of Distance on Tourism in Hong Kong: A Comparison of Short Haul and Long Haul Visitors.” *Asia Pacific Journal of Tourism Research* 13 (2): 101–11. https://doi.org/10.1080/10941660802048332.

Farber, Steven, Morton O’Kelly, Harvey J. Miller, and Tijs Neutens. 2015. “Measuring Segregation Using Patterns of Daily Travel Behavior: A Social Interaction Based Model of Exposure.” *Journal of Transport Geography* 49: 26–38. https://doi.org/10.1016/j.jtrangeo.2015.10.009.

Farber, Steven, Antonio Páez, and Catherine Morency. 2012. “Activity Spaces and the Measurement of Clustering and Exposure: A Case Study of Linguistic Groups in Montreal.” *Environment and Planning A* 44 (2): 315–32. https://doi.org/10.1068/a44203.

Feitosa, F. F., G. Câmara, A. M. V. Monteiro, T. Koschitzki, and M. P. S. Silva. 2007. “Global and Local Spatial Indices of Urban Segregation.” *International Journal of Geographical Information Science* 21 (3): 299–323. https://doi.org/10.1080/13658810600911903.

Ferrari, Laura, Alberto Rosi, Marco Mamei, and Franco Zambonelli. 2011. “Extracting Urban Patterns from Location-Based Social Networks.” In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks - LBSN ’11*, 1. New York, New York, USA: ACM Press. https://doi.org/10.1145/2063212.2063226.

Frias-Martinez, Vanessa, and Enrique Frias-Martinez. 2014. “Spectral Clustering for Sensing Urban Land Use Using Twitter Activity.” *Engineering Applications of Artificial Intelligence* 35 (October): 237–45. https://doi.org/10.1016/j.engappai.2014.06.019.

Gabrielli, Lorenzo, Barbara Furletti, Roberto Trasarti, Fosca Giannotti, and Dino Pedreschi. 2015. “City Users’ Classification with Mobile Phone Data.” In *2015 IEEE International Conference on Big Data (Big Data)*, 1007–12. IEEE. https://doi.org/10.1109/BigData.2015.7363852.

Galster, George C, and Sean P Killen. 1995. “The Geography of Metropolitan Opportunity: A

Reconnaissance and Conceptual Framework." *Housing Policy Debate* 6 (1): 7–43. https://doi.org/10.1080/10511482.1995.9521180.

Gao, Song, Krzysztof Janowicz, Daniel R. Montello, Yingjie Hu, Jiue-An Yang, Grant McKenzie, Yiting Ju, Li Gong, Benjamin Adams, and Bo Yan. 2017. "A Data-Synthesis-Driven Method for Detecting and Extracting Vague Cognitive Regions." *International Journal of Geographical Information Science* 00 (00): 1–27. https://doi.org/10.1080/13658816.2016.1273357.

Gao, Song, Yaoli Wang, Yong Gao, and Yu Liu. 2013. "Understanding Urban Traffic-Flow Characteristics: A Rethinking of Betweenness Centrality." *Environment and Planning B: Planning and Design* 40 (1): 135–53. https://doi.org/10.1068/b38141.

Gao, Song, JA Yang, Bo Yan, Yingjie Hu, Krzysztof Janowicz, and G McKenzie. 2014. "Detecting Origin-Destination Mobility Flows From Geotagged Tweets in Greater Los Angeles Area." *Geog.Ucsb.Edu*, 0–4. http://www.geog.ucsb.edu/~sgao/papers/2014_GIScience_EA_DetectingODTripsUsingGeoTweets.pdf.

García-Palomares, Juan Carlos, Javier Gutiérrez, and Carmen Mínguez. 2015. "Identification of Tourist Hot Spots Based on Social Networks: A Comparative Analysis of European Metropolises Using Photo-Sharing Services and GIS." *Applied Geography* 63 (September): 408–17. https://doi.org/10.1016/j.apgeog.2015.08.002.

Girardin, Fabien, Francesco Calabrese, Filippo Dal Fiore, Carlo Ratti, and Josep Blat. 2008. "Digital Footprinting: Uncovering Tourists with User-Generated Content." *IEEE Pervasive Computing* 7 (4): 36–43. https://doi.org/10.1109/MPRV.2008.71.

Glaser, Susan. 2016. "Greater Cleveland Attracted 17.6 Million Visitors in 2015, Another Record." 2016. http://www.cleveland.com/travel/index.ssf/2016/09/greater_cleveland_attracted_17.html.

González, Marta C, César a Hidalgo, and Albert-László Barabási. 2008. "Understanding Individual Human Mobility Patterns." *Nature* 453 (7196): 779–82. https://doi.org/10.1038/nature06958.

Goodchild, Michael F. 2007. “Citizens as Sensors: The World of Volunteered Geography.” *GeoJournal*. https://doi.org/10.1007/s10708-007-9111-y.

Guo, Diansheng, Xi Zhu, Hai Jin, Peng Gao, and Clio Andris. 2012. “Discovering Spatial Patterns in Origin-Destination Mobility Data.” *Transactions in GIS* 16 (3): 411–29. https://doi.org/10.1111/j.1467-9671.2012.01344.x.

Han, Shanshan, Fu Ren, Chao Wu, Ying Chen, Qingyun Du, and Xinyue Ye. 2018. “Using the TensorFlow Deep Neural Network to Classify Mainland China Visitor Behaviours in Hong Kong from Check-in Data.” *ISPRS International Journal of Geo-Information* 7 (4): 158. https://doi.org/10.3390/ijgi7040158.

Hargittai, Eszter, and Eden Litt. 2011. “The Tweet Smell of Celebrity Success: Explaining Variation in Twitter Adoption among a Diverse Group of Young Adults.” *New Media & Society* 13 (5): 824–42. https://doi.org/10.1177/1461444811405805.

Hawelka, Bartosz, Izabela Sitko, Euro Beinat, Stanislav Sobolevsky, Pavlos Kazakopoulos, and Carlo Ratti. 2014. “Geo-Located Twitter as Proxy for Global Mobility Patterns.” *Cartography and Geographic Information Science* 41 (3): 260–71. https://doi.org/10.1080/15230406.2014.890072.

Hecht, Brent, and Monica Stephens. 2014. “A Tale of Cities: Urban Biases in Volunteered Geographic Information.” *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*, 197–205.

Hong, Xin, and Xinyue Ye. 2018. “Exploring the Influence of Land Cover on Weight Loss Awareness.” *GeoJournal* 83 (5): 935–47. https://doi.org/10.1007/s10708-017-9806-7.

Horner, Mark W., and Jessica N. Mefford. 2007. “Investigating Urban Spatial Mismatch Using Job-Housing Indicators to Model Home-Work Separation.” *Environment and Planning A* 39 (6): 1420–40. https://doi.org/10.1068/a37443.

Horton, Frank E, and David R Reynolds. 1971. “Effects of Urban Spatial Structure on Individual Behavior.” *Economic Geography* 47 (1): 36. https://doi.org/10.2307/143224.

Hu, Lingqian, Zhenlong Li, and Xinyue Ye. 2020. “Delineating and Modeling Activity Space Using Geotagged Social Media Data.” *Cartography and Geographic Information Science*

47 (3): 277–88. https://doi.org/10.1080/15230406.2019.1705187.

Hu, Yingjie, Song Gao, Krzysztof Janowicz, Bailang Yu, Wenwen Li, and Sathya Prasad. 2015. "Extracting and Understanding Urban Areas of Interest Using Geotagged Photos." *Computers, Environment and Urban Systems* 54 (November): 240–54. https://doi.org/10.1016/j.compenvurbsys.2015.09.001.

Huang, Qunying. 2016. "Mining Online Footprints to Predict User's next Location." *International Journal of Geographical Information Science* 8816 (August): 1–19. https://doi.org/10.1080/13658816.2016.1209506.

Huang, Qunying, and David W. S. Wong. 2016. "Activity Patterns, Socioeconomic Status and Urban Spatial Structure: What Can Social Media Data Tell Us?" *International Journal of Geographical Information Science* 8816 (February): 1–26. https://doi.org/10.1080/13658816.2016.1145225.

Huang, Qunying, and David W.S. Wong. 2015a. "Modeling and Visualizing Regular Human Mobility Patterns with Uncertainty: An Example Using Twitter Data." *Annals of the Association of American Geographers*. https://doi.org/10.1080/00045608.2015.1081120.

Huang, Qunying, and David W S Wong. 2015b. "Modeling and Visualizing Regular Human Mobility Patterns with Uncertainty : An Example Using Twitter Data." *Annals of the Association of American Geographers* 105(6) (November): 1179–97. https://doi.org/10.1080/00045608.2015.1081120.

Hughes, Holly L. 1993. "Metropolitan Structure and the Suburban Hierarchy." *American Sociological Review* 58 (3): 417–33.

Iqbal, Md Shahadat, Charisma F. Choudhury, Pu Wang, and Marta C. González. 2014. "Development of Origin–Destination Matrices Using Mobile Phone Call Data." *Transportation Research Part C: Emerging Technologies* 40 (March): 63–74. https://doi.org/10.1016/j.trc.2014.01.002.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. Springer Texts in Statistics. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4614-7138-7.

Järv, Olle, Rein Ahas, and Frank 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: 122–35. https://doi.org/10.1016/j.trc.2013.11.003.

Jones, Miranda R, Ana V. Diez-Roux, Anjum Hajat, Kiarri N Kershaw, Marie S. O’Neill, Eliseo Guallar, Wendy S. Post, Joel D. Kaufman, and Ana Navas-Acien. 2014. “Race/Ethnicity, Residential Segregation, and Exposure to Ambient Air Pollution: The Multi-Ethnic Study of Atherosclerosis (MESA).” *American Journal of Public Health* 104 (11): 2130–37. https://doi.org/10.2105/AJPH.2014.302135.

Jurdak, Raja, Kun Zhao, Jiajun Liu, Maurice AbouJaoude, Mark Cameron, and David Newth. 2015. “Understanding Human Mobility from Twitter.” Edited by Ye Wu. *PLOS ONE* 10 (7): e0131469. https://doi.org/10.1371/journal.pone.0131469.

Kang, Chaogui, Xiujun Ma, Daoqin Tong, and Yu Liu. 2012. “Intra-Urban Human Mobility Patterns: An Urban Morphology Perspective.” *Physica A: Statistical Mechanics and Its Applications* 391 (4): 1702–17. https://doi.org/10.1016/j.physa.2011.11.005.

Kerkvliet, Joe, and Clifford Nowell. 1999. “Heterogeneous Visitors and the Spatial Limits of the Travel Cost Model.” *Journal of Leisure Research* 31 (4): 404–19. https://doi.org/10.1080/00222216.1999.11949874.

Kwan, Mei-po. 2016. “Algorithmic Geographies: Big Data, Algorithmic Uncertainty, and the Production of Geographic Knowledge.” *Annals of the American Association of Geographers* 106 (2): 274–82. https://doi.org/10.1080/00045608.2015.1117937.

Kwan, Mei-Po. 1999. “Gender, the Home-Work Link, and Space-Time Patterns of Nonemployment Activities.” *Economic Geography* 75 (4): 370–94. https://doi.org/10.2307/144477.

———. 2008. “From Oral Histories to Visual Narratives: Re-Presenting the Post-September 11 Experiences of the Muslim Women in the USA.” *Social & Cultural Geography* 9 (6): 653–69. https://doi.org/10.1080/14649360802292462.

———. 2015. “Beyond Space (As We Knew It): Toward Temporally Integrated Geographies of

Segregation, Health, and Accessibility." In *Space-Time Integration in Geography and GIScience*, 5608:39–51. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-017-9205-9_4.

Kwan, Mei Po. 2009. "From Place-Based to People-Based Exposure Measures." *Social Science and Medicine* 69 (9): 1311–13. https://doi.org/10.1016/j.socscimed.2009.07.013.

Lambert, Dayton M., Jason P. Brown, and Raymond J.G.M. Florax. 2010. "A Two-Step Estimator for a Spatial Lag Model of Counts: Theory, Small Sample Performance and an Application." *Regional Science and Urban Economics* 40 (4): 241–52. https://doi.org/10.1016/j.regsciurbeco.2010.04.001.

Lee, Jae Hyun, Song Gao, and Konstadinos G. Goulias. 2015. "Can Twitter Data Be Used to Validate Travel Demand Models?" In *In 14th International Conference on Travel Behaviour Research*. Windsor, UK. https://www.researchgate.net/publication/282171236_Can_Twitter_data_be_used_to_validate_travel_demand_models.

Lee, Jay, and Shengwen Li. 2017. "Extending Moran's Index for Measuring Spatiotemporal Clustering of Geographic Events." *Geographical Analysis* 49 (1): 36–57. https://doi.org/10.1111/gean.12106.

Lenormand, Maxime, Miguel Picornell, Oliva G. Cantú-Ros, Antònia Tugores, Thomas Louail, Ricardo Herranz, Marc Barthelemy, Enrique Frías-Martínez, and José J. Ramasco. 2014. "Cross-Checking Different Sources of Mobility Information." Edited by Yamir Moreno. *PLoS ONE* 9 (8): e105184. https://doi.org/10.1371/journal.pone.0105184.

LeSage, James P., Manfred M. Fischer, and Thomas Scherngell. 2007. "Knowledge Spillovers across Europe: Evidence from a Poisson Spatial Interaction Model with Spatial Effects." *Papers in Regional Science* 86 (3): 393–421. https://doi.org/10.1111/j.1435-5957.2007.00125.x.

LeSage, James P., and Robert Kelley Pace. 2008. "Spatial Econometric Modeling of Origin-Destination Flows." *Journal of Regional Science* 48 (5): 941–67. https://doi.org/10.1111/j.1467-9787.2008.00573.x.

———. 2009. *Introduction to Spatial Econometrics*. CRC Press.

Leung, Daniel, Rob Law, Hubert van Hoof, and Dimitrios Buhalis. 2013. "Social Media in Tourism and Hospitality: A Literature Review." *Journal of Travel & Tourism Marketing* 30 (1–2): 3–22. https://doi.org/10.1080/10548408.2013.750919.

Li, Dongying, Xiaolu Zhou, and Mingshu Wang. 2018. "Analyzing and Visualizing the Spatial Interactions between Tourists and Locals: A Flickr Study in Ten US Cities." *Cities* 74 (January): 249–58. https://doi.org/10.1016/j.cities.2017.12.012.

Li, Linna, Michael F. Goodchild, and Bo Xu. 2013. "Spatial, Temporal, and Socioeconomic Patterns in the Use of Twitter and Flickr." *Cartography and Geographic Information Science* 40 (2): 61–77. https://doi.org/10.1080/15230406.2013.777139.

Liao, Chuan, Daniel Brown, Ding Fei, Xuewen Long, Dan Chen, and Shengquan Che. 2018. "Big Data-enabled Social Sensing in Spatial Analysis: Potentials and Pitfalls." *Transactions in GIS* 22 (6): 1351–71. https://doi.org/10.1111/tgis.12483.

Liu, Qingsong, Zheye Wang, and Xinyue Ye. 2018. "Comparing Mobility Patterns between Residents and Visitors Using Geo-tagged Social Media Data." *Transactions in GIS* 22 (6): 1372–89. https://doi.org/10.1111/tgis.12478.

Liu, X, C Kang, L Gong, and Y Liu. 2016. "Incorporating Spatial Interaction Patterns in Classifying and Understanding Urban Land Use." *International Journal of Geographical Information Science* 30 (2): 334–50. https://doi.org/10.1080/13658816.2015.1086923.

Liu, Yu, Zhengwei Sui, Chaogui Kang, and Yong Gao. 2014. "Uncovering Patterns of Inter-Urban Trip and Spatial Interaction from Social Media Check-In Data." Edited by Peter Csermely. *PLoS ONE* 9 (1): e86026. https://doi.org/10.1371/journal.pone.0086026.

Long, Jed a., and Trisalyn a. Nelson. 2012. "A Review of Quantitative Methods for Movement Data." *International Journal of Geographical Information Science* 27 (2): 1–27. https://doi.org/10.1080/13658816.2012.682578.

Long, Ying, Xingjian Liu, Jiangping Zhou, and Yanwei Chai. 2016. "Early Birds, Night Owls, and Tireless/Recurring Itinerants: An Exploratory Analysis of Extreme Transit Behaviors in Beijing, China." *Habitat International* 57 (51408039): 223–32.

https://doi.org/10.1016/j.habitatint.2016.08.004.

Luo, Feixiong, Guofeng Cao, Kevin Mulligan, and Xiang Li. 2016. “Explore Spatiotemporal and Demographic Characteristics of Human Mobility via Twitter: A Case Study of Chicago.” *Applied Geography* 70: 11–25. https://doi.org/10.1016/j.apgeog.2016.03.001.

Massey, Douglas S., and Nancy A. Denton. 1988. “The Dimensions of Residential Segregation.” *Social Forces* 67 (2): 281. https://doi.org/10.2307/2579183.

Mevik, Bjørn-Helge, and Ron Wehrens. 2007. “The Pls Package: Principal Component and Partial Least Squares Regression in R.” *Journal of Statistical Software* 18 (2). https://doi.org/10.18637/jss.v018.i02.

Miah, Shah Jahan, Huy Quan Vu, John Gammack, and Michael McGrath. 2017. “A Big Data Analytics Method for Tourist Behaviour Analysis.” *Information & Management* 54 (6): 771–85. https://doi.org/10.1016/j.im.2016.11.011.

Mislove, Alan, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosenquist. 2011. “Understanding the Demographics of Twitter Users.” *Int’l AAAI Conference on Weblogs and Social Media (ICWSM)* 5 (1): 554–57.

Morgan, Barrie S. 1983. “A Distance-Decay Based Interaction Index to Measure Residential Segregation.” *Area* 15 (3): 211–17.

Morril, Richard L. 1991. “On the Measure of Spatial Segregation.” *Geography Research Forum* 11: 25–36. http://raphael.geography.ad.bgu.ac.il/ojs/index.php/GRF/article/viewFile/91/87.

Naess, P. 2000. “Urban Structures and Travel Behaviour. Experiences from Empirical Research in Norway and Denmark.” *Land Use and Travel Behaviour*, 1–24. http://www.ejtir.tbm.tudelft.nl/issues/2003_02/pdf/2003_02_03.pdf.

Nengsih, Titin Agustin, Frédéric Bertrand, Myriam Maumy-Bertrand, and Nicolas Meyer. 2019. “Determining the Number of Components in PLS Regression on Incomplete Data Set.” *Statistical Applications in Genetics and Molecular Biology* 18 (6). https://doi.org/10.1515/sagmb-2018-0059.

Nijman, Jan, and Yehua Dennis Wei. 2020. “Urban Inequalities in the 21st Century Economy.”

Applied Geography 117 (April): 102188. https://doi.org/10.1016/j.apgeog.2020.102188.

Orsi, Francesco, and Davide Geneletti. 2013. “Using Geotagged Photographs and GIS Analysis to Estimate Visitor Flows in Natural Areas.” *Journal for Nature Conservation* 21 (5): 359–68. https://doi.org/10.1016/j.jnc.2013.03.001.

Park, Yoo Min, and Mei-po Kwan. 2018. “Beyond Residential Segregation: A Spatiotemporal Approach to Examining Multi-Contextual Segregation.” *Computers, Environment and Urban Systems* 71 (May): 98–108. https://doi.org/10.1016/j.compenvurbsys.2018.05.001.

Park, Yoo Min, and Mei Po Kwan. 2017a. “Individual Exposure Estimates May Be Erroneous When Spatiotemporal Variability of Air Pollution and Human Mobility Are Ignored.” *Health and Place* 43 (October 2016): 85–94. https://doi.org/10.1016/j.healthplace.2016.10.002.

———. 2017b. “Multi-Contextual Segregation and Environmental Justice Research: Toward Fine-Scale Spatiotemporal Approaches.” *International Journal of Environmental Research and Public Health* 14 (10). https://doi.org/10.3390/ijerph14101205.

Pei, Tao, Stanislav Sobolevsky, Carlo Ratti, Shih-Lung Shaw, Ting Li, and Chenghu Zhou. 2014. “A New Insight into Land Use Classification Based on Aggregated Mobile Phone Data.” *International Journal of Geographical Information Science* 28 (9): 1988–2007. https://doi.org/10.1080/13658816.2014.913794.

Phillips, Nolan E., Brian L. Levy, Robert J. Sampson, Mario L. Small, and Ryan Q. Wang. 2019. “The Social Integration of American Cities: Network Measures of Connectedness Based on Everyday Mobility Across Neighborhoods.” *Sociological Methods and Research* 0049124119. https://doi.org/10.1177/0049124119852386.

Phillips, WooMi J., and SooCheong Jang. 2010. “Destination Image Differences between Visitors and Non-Visitors: A Case of New York City.” *International Journal of Tourism Research* 12 (5): 642–45. https://doi.org/10.1002/jtr.776.

Reardon, Sean F., and David O’Sullivan. 2004. “Measures of Spatial Segregation.” *Sociological Methodology* 34 (1): 121–62. https://doi.org/10.1111/j.0081-1750.2004.00150.x.

Sampson, Robert J, and Brian L Levy. 2020. “Beyond Residential Segregation: Mobility-Based

Connectedness and Rates of Violence in Large Cities." *Race and Social Problems* 12 (1): 77–86. https://doi.org/10.1007/s12552-019-09273-0.

Schönfelder, Stefan, and Kay W Axhausen. 2003. "Activity Spaces: Measures of Social Exclusion?" *Transport Policy* 10 (4): 273–86. https://doi.org/10.1016/j.tranpol.2003.07.002.

Shaw, Shih-Lung, Ming-Hsiang Tsou, and Xinyue Ye. 2016. "Editorial: Human Dynamics in the Mobile and Big Data Era." *International Journal of Geographical Information Science* 30 (9): 1687–93. https://doi.org/10.1080/13658816.2016.1164317.

Shelton, Taylor, Ate Poorthuis, Mark Graham, and Matthew Zook. 2014. "Mapping the Data Shadows of Hurricane Sandy: Uncovering the Sociospatial Dimensions of 'Big Data.'" *Geoforum* 52 (March): 167–79. https://doi.org/10.1016/j.geoforum.2014.01.006.

Shelton, Taylor, Ate Poorthuis, and Matthew Zook. 2015. "Social Media and the City: Rethinking Urban Socio-Spatial Inequality Using User-Generated Geographic Information." *Landscape and Urban Planning* 142 (October): 198–211. https://doi.org/10.1016/j.landurbplan.2015.02.020.

Shi, Li, Guanghua Chi, Xi Liu, and Yu Liu. 2015. "Human Mobility Patterns in Different Communities: A Mobile Phone Data-Based Social Network Approach." *Annals of GIS* 21 (1): 15–26. https://doi.org/10.1080/19475683.2014.992372.

Shoval, Noam, and Michal Isaacson. 2007. "Tracking Tourists in the Digital Age." *Annals of Tourism Research* 34 (1): 141–59. https://doi.org/10.1016/j.annals.2006.07.007.

Sikkink, David, and Michael O. Emerson. 2008. "School Choice and Racial Segregation in US Schools: The Role of Parents' Education." *Ethnic and Racial Studies* 31 (2): 267–93. https://doi.org/10.1080/01419870701337650.

Silm, Siiri, and Rein Ahas. 2014a. "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 (3): 542–59. https://doi.org/10.1080/00045608.2014.892362.

———. 2014b. "The Temporal Variation of Ethnic Segregation in a City: Evidence from a Mobile Phone Use Dataset." *Social Science Research* 47 (September): 30–43. https://doi.org/10.1016/j.ssresearch.2014.03.011.

Song, Chaoming, Zehui Qu, Nicholas Blumm, and A.-L. Barabasi. 2010. “Limits of Predictability in Human Mobility.” *Science* 327 (5968): 1018–21. https://doi.org/10.1126/science.1177170.

Steiger, Enrico, René Westerholt, Bernd Resch, and Alexander Zipf. 2015. “Twitter as an Indicator for Whereabouts of People? Correlating Twitter with UK Census Data.” *Computers, Environment and Urban Systems* 54 (November): 255–65. https://doi.org/10.1016/j.compenvurbsys.2015.09.007.

Straumann, Ralph K., Arzu Çöltekin, and Gennady Andrienko. 2014. “Towards (Re)Constructing Narratives from Georeferenced Photographs through Visual Analytics.” *The Cartographic Journal* 51 (2): 152–65. https://doi.org/10.1179/1743277414Y.0000000079.

Sultana, Selima. 2005. “Racial Variations in Males’ Commuting Times in Atlanta: What Does the Evidence Suggest?” *Professional Geographer* 57 (1): 66–82. https://doi.org/10.1111/j.0033-0124.2005.00460.x.

Tang, Jinjun, Fang Liu, Yinhai Wang, and Hua Wang. 2015. “Uncovering Urban Human Mobility from Large Scale Taxi GPS Data.” *Physica A: Statistical Mechanics and Its Applications* 438 (December 2017): 140–53. https://doi.org/10.1016/j.physa.2015.06.032.

Toole, Jameson L., Carlos Herrera-Yaqüe, Christian M. Schneider, and Marta C. González. 2015. “Coupling Human Mobility and Social Ties.” *Journal of The Royal Society Interface* 12 (105): 20141128. https://doi.org/10.1098/rsif.2014.1128.

Tsou, Ming-Hsiang, Hao Zhang, Atsushi Nara, and Su Yeon Han. 2018. “Estimating Hourly Population Distribution Change at High Spatiotemporal Resolution in Urban Areas Using Geo-Tagged Tweets, Land Use Data, and Dasymetric Maps.” *Journal of Statistical Software* 18 (2). http://arxiv.org/abs/1810.06554.

Turner, Lena Magnusson, and Terje Wessel. 2013. “Upwards, Outwards and Westwards: Relocation of Ethnic Minority Groups in the Oslo Region.” *Geografiska Annaler: Series B, Human Geography* 95 (1): 1–16. https://doi.org/10.1111/geob.12006.

Vu, Huy Quan, Gang Li, Rob Law, and Ben Haobin Ye. 2015. “Exploring the Travel Behaviors

of Inbound Tourists to Hong Kong Using Geotagged Photos." *Tourism Management* 46 (February): 222–32. https://doi.org/10.1016/j.tourman.2014.07.003.

Wang, Donggen, Fei Li, and Yanwei Chai. 2012. "Activity Spaces and Sociospatial Segregation in Beijing." *Urban Geography* 33 (2): 256–77. https://doi.org/10.2747/0272-3638.33.2.256.

Wang, Donggen, and Meng Zhou. 2017. "The Built Environment and Travel Behavior in Urban China: A Literature Review." *Transportation Research Part D: Transport and Environment* 52 (May): 574–85. https://doi.org/10.1016/j.trd.2016.10.031.

Wang, Qi, Nolan Edward Phillips, Mario L. Small, and Robert J. Sampson. 2018. "Urban Mobility and Neighborhood Isolation in America's 50 Largest Cities." *Proceedings of the National Academy of Sciences of the United States of America* 115 (30): 7735–40. https://doi.org/10.1073/pnas.1802537115.

Wang, Qi, and John E. Taylor. 2014. "Quantifying Human Mobility Perturbation and Resilience in Hurricane Sandy." *PLoS ONE* 9 (11): 1–5. https://doi.org/10.1371/journal.pone.0112608.

Wang, Zheye, Nina S.N. Lam, Nick Obradovich, and Xinyue Ye. 2019. "Are Vulnerable Communities Digitally Left behind in Social Responses to Natural Disasters? An Evidence from Hurricane Sandy with Twitter Data." *Applied Geography* 108 (February): 1–8. https://doi.org/10.1016/j.apgeog.2019.05.001.

Wang, Zheye, Xinyue Ye, Jay Lee, Xiaomeng Chang, Haimeng Liu, and Qingquan Li. 2018. "A Spatial Econometric Modeling of Online Social Interactions Using Microblogs." *Computers, Environment and Urban Systems* 70 (September 2017): 53–58. https://doi.org/10.1016/j.compenvurbsys.2018.02.001.

Warf, Barney, and Brian Holly. 1997. "The Rise and Fall and Rise of Cleveland." *The ANNALS of the American Academy of Political and Social Science* 551 (1): 208–21. https://doi.org/10.1177/0002716297551001015.

Weaver, David B., Anna Kwek, and Ying Wang. 2017. "Cultural Connectedness and Visitor Segmentation in Diaspora Chinese Tourism." *Tourism Management* 63 (December): 302–14. https://doi.org/10.1016/j.tourman.2017.06.028.

White, Kellee, and Luisa N Borrell. 2011. "Racial/Ethnic Residential Segregation: Framing the

Context of Health Risk and Health Disparities." *Health & Place* 17 (2): 438–48. https://doi.org/https://doi.org/10.1016/j.healthplace.2010.12.002.

White, Michael J. 1983. "The Measurement of Spatial Segregation." *American Journal of Sociology* 88 (5): 1008–18. http://www.jstor.org/stable/27794.

Williams, David R., and Chiquita Collins. 2001. "Racial Residential Segregation: A Fundamental Cause of Racial Disparities in Health." *Public Health Reports* 116 (5): 404–16. https://doi.org/10.1016/S0033-3549(04)50068-7.

Wojcik, Stefan, and Adam Hughes. 2019. "Sizing Up Twitter Users." 2019. https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/.

Wong, David W. S. 1993. "Spatial Indices of Segregation." *Urban Studies* 30 (3): 559–72. http://www.stata.com/meeting/italy09/italy09_pisati.pdf.

———. 1998. "Measuring Multiethnic Spatial Segregation." *Urban Geography* 19 (1): 77–87. https://doi.org/10.2747/0272-3638.19.1.77.

Wong, David W. S., and Shih-Lung Shaw. 2011. "Measuring Segregation: An Activity Space Approach." *Journal of Geographical Systems* 13 (2): 127–45. https://doi.org/10.1007/s10109-010-0112-x.

Wong, David W.S. 2002. "Modeling Local Segregation: A Spatial Interaction Approach." *Geographical and Environmental Modelling* 6 (1): 81–97. https://doi.org/10.1080/13615930220127305.

Xiang, Zheng, and Ulrike Gretzel. 2010. "Role of Social Media in Online Travel Information Search." *Tourism Management* 31 (2): 179–88. https://doi.org/10.1016/j.tourman.2009.02.016.

Xu, Yang, Alexander Belyi, Iva Bojic, and Carlo Ratti. 2017. "How Friends Share Urban Space: An Exploratory Spatiotemporal Analysis Using Mobile Phone Data." *Transactions in GIS* 21 (3): 468–87. https://doi.org/10.1111/tgis.12285.

Xu, Yang, Alexander Belyi, Paolo Santi, and Carlo Ratti. 2019. "Quantifying Segregation in an Integrated Urban Physical-Social Space." *Journal of The Royal Society Interface* 16 (160):

20190536. https://doi.org/10.1098/rsif.2019.0536.

Yang, Gege, Ci Song, Hua Shu, Jia Zhang, Tao Pei, and Chenghu Zhou. 2016. “Assessing Patient Bypass Behavior Using Taxi Trip Origin–Destination (OD) Data.” *ISPRS International Journal of Geo-Information* 5 (9): 157. https://doi.org/10.3390/ijgi5090157.

Yin, Junjun, Aiman Soliman, Dandong Yin, and Shaowen Wang. 2017. “Depicting Urban Boundaries from a Mobility Network of Spatial Interactions: A Case Study of Great Britain with Geo-Located Twitter Data.” *International Journal of Geographical Information Science* 31 (7): 1293–1313. https://doi.org/10.1080/13658816.2017.1282615.

Yip, Ngai Ming, Ray Forrest, and Shi Xian. 2016. “Exploring Segregation and Mobilities: Application of an Activity Tracking App on Mobile Phone.” *Cities* 59: 156–63. https://doi.org/10.1016/j.cities.2016.02.003.

Zeng, Benxiang, and Rolf Gerritsen. 2014. “What Do We Know about Social Media in Tourism? A Review.” *Tourism Management Perspectives* 10: 27–36. https://doi.org/10.1016/j.tmp.2014.01.001.

Zhao, Ziliang, Shih Lung Shaw, Yang Xu, Feng Lu, Jie Chen, and Ling Yin. 2016. “Understanding the Bias of Call Detail Records in Human Mobility Research.” *International Journal of Geographical Information Science* 30 (9): 1738–62. https://doi.org/10.1080/13658816.2015.1137298.

Zheng, Yu U, Licia Capra, Ouri Wolfson, and Hai A I Yang. 2014. “Urban Computing : Concepts , Methodologies , and Applications.” *ACM Transactions on Intelligent Systems and Technology* 5 (3): 1–55. https://doi.org/10.1145/2629592.

Zhou, Xiaolu, Chen Xu, and Brandon Kimmons. 2015. “Detecting Tourism Destinations Using Scalable Geospatial Analysis Based on Cloud Computing Platform.” *Computers, Environment and Urban Systems* 54 (November): 144–53. https://doi.org/10.1016/j.compenvurbsys.2015.07.006.

Zhu, Di, Ninghua Wang, Lun Wu, and Yu Liu. 2017. “Street as a Big Geo-Data Assembly and Analysis Unit in Urban Studies: A Case Study Using Beijing Taxi Data.” *Applied Geography* 86 (September): 152–64. https://doi.org/10.1016/j.apgeog.2017.07.001.