12-2015

Cell Towers as Urban Sensors: Understanding the Strengths and Limitations of Mobile Phone Location Data

Ziliang Zhao
University of Tennessee - Knoxville, zzhao7@vols.utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Ziliang Zhao entitled "Cell Towers as Urban Sensors: Understanding the Strengths and Limitations of Mobile Phone Location Data." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Geography.

Shih-Lung Shaw, Major Professor

We have read this dissertation and recommend its acceptance:

Bruce Ralston, Hyun Kim, Dali Wang, Lee Han

Accepted for the Council:
Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Cell Towers as Urban Sensors: Understanding the Strengths and Limitations of Mobile Phone Location Data

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Ziliang Zhao
December 2015

DEDICATION

This dissertation is dedicated to my parents and my grandparents, for their continual and comprehensive support since I was born. I also dedicate this dissertation to my wife, Qingmei, who sacrificed her own career to take good care of me. Finally, this dissertation is dedicated to my upcoming son, Roger, a wonderful gift that comes with the accomplishment of my PhD research.

ACKNOWLEDGEMENTS

I am very grateful to my advisor, Dr. Shih-Lung Shaw. Dr. Shaw has provided me with consistent support since he brought me in as a master student in Fall 2009. His guidance has lifted me to a higher level in terms of independent thinking and research, and changed me from a pure technician to a person who has the capability to address real-world problems. Besides academic guidance, from Dr. Shaw I have also learnt the way of doing things correctly and professionally. This will certainly be invaluable for the rest of my life.

Many thanks go to Dr. Bruce Ralston. Dr. Ralston led me to the world of the exciting GIS technologies, which laid down the foundation for my PhD research and future career. Dr. Ralston always stays hungry to learn new things even after his retirement. Such spirit encourages me to never stop learning. In addition, his optimistic attitude makes me more determined about doing things I am passionate about.

I am also grateful to the assistance from Dr. Hyun Kim, Dr. Dali Wang, and Dr. Lee Han. They kindly served in my dissertation committee and offered useful suggestions for improving my research.

Finally, I hope to express my appreciation to other fellow students/friends, including, but not limited to, Ling Yin, Yitu Xu, Jie Chen, Yang Xu, and Jiaoli Chen. It has been a great pleasure to learn from all of you.

ABSTRACT

Understanding urban dynamics and human mobility patterns not only benefits a wide range of real-world applications (e.g., business site selection, public transit planning), but also helps address many urgent issues caused by the rapid urbanization processes (e.g., population explosion, congestion, pollution). In the past few years, given the pervasive usage of mobile devices, call detail records collected by mobile network operators has been widely used in urban dynamics and human mobility studies. However, the derived knowledge might be strongly biased due to the uneven distribution of people's phone communication activities in space and time.

This dissertation research applies different analytical methods to better understand human activity and urban environment, as well as their interactions, mainly based on a new type of data source: actively tracked mobile phone location data. In particular, this dissertation research achieves three main research objectives. First, this research develops visualization and analysis approaches to uncover hidden urban dynamics patterns from actively tracked mobile phone location data. Second, this research designs quantitative methods to evaluate the representativeness issue of call detail record data. Third, this research develops an appropriate approach to evaluate the performance of different types of tracking data in urban dynamics research.

The major contributions of this dissertation research include: 1) uncovering the dynamics of stay/move activities and distance decay effects, and the changing human mobility patterns based on several mobility indicators derived from actively tracked mobile phone location data; 2) taking the first step to evaluate the representativeness and effectiveness of call detail record and revealing its bias in human mobility research; and 3) extracting and comparing urban-level population movement patterns derived from three different types of tracking data as well as their pros and cons in urban population movement analysis.

TABLE OF CONTENTS

Chapter 1 Introduction 1
1.1 Research Background and Research Questions 2
1.2 Organization of the Dissertation 6
References 8
Chapter 2 Cell Towers as Sensors: Uncovering the Pulse of a City Based on Actively Tracked Mobile Phone Location Data 11
Abstract 12
2.1 Introduction 12
2.2 Related Research 15
2.2.1 Geovisualization of human mobility 15
2.2.2 Mobile phone location data and urban dynamics 16
2.3 Dataset 17
2.3.1 Study area 17
2.3.2 Dataset 18
2.4 Examine urban dynamics with *STEAM* 19
2.4.1 Stay/move dynamics 20
2.4.2 Distance decay dynamics 21
2.5 Uncover urban dynamics using mobility time series 22
2.5.1 Method 22
2.5.2 Results 24
2.6 Conclusions 31
References 34
Chapter 3 Understanding the Bias of Call Detail Records in Human Mobility Research 38
Abstract 39
3.1 Introduction 39
3.2 Relevant research 41
3.2.1 CDRs and human mobility 41
3.2.2 CDRs and urban dynamics 42
3.2.3 Uncertainty issue 43
3.3 Data 44
3.3.1 Area of study 44
3.3.2 Dataset 44
3.3.3 Data processing 45
3.4 Individual human mobility 48
3.4.1 Total travel distance 49
3.4.2 Radius of gyration 51
3.4.3 Movement entropy 53
3.5 Collective human mobility 56
3.5.1 Distance decay effect 56
3.5.2 Community detection 57
3.6 Conclusions and Discussions 63

References 65
Chapter 4 Extract and compare generalized population movement patterns derived from different tracking datasets using a revised hierarchical clustering algorithm 69
Abstract 70
4.1 Introduction 70
4.2 Relevant research 72
4.2.1 Big tracking data and urban dynamics 72
4.2.2 Flow data aggregation and clustering 73
4.3 Method 75
4.3.1 A revised hierarchical clustering algorithm 75
4.3.2 Sensitivity analysis of k 77
4.3.3 Two-step hierarchical clustering 78
4.4 Compare generalized urban population movement patterns in Shenzhen 81
4.4.1 Study area and the datasets 81
4.4.2 Results 82
4.4.3 Discussions 87
4.5 Conclusions and future work 89
References 91
Chapter 5 Conclusions 95
5.1 Summary 96
5.2 Potential Applications 97
5.2.1 STEAM and geovisualization 97
5.2.2 Variation pattern of urban dynamics 98
5.2.3 Hierarchical flow clustering 98
5.3 Future work 100
5.3.1 Validation of “patterns” 100
5.3.2 The gap between “patterns” and “processes” 100
5.3.3 Data fusion and urban dynamics research 101
References 103
VITA 104

LIST OF TABLES

Table 3.1 Summary of event codes. 46
Table 3.2 Summary of four subscriber classes divided by CDR ratio. 49
Table 3.3 Correlation between total travel distance (complete group) and average total travel distance (CDR group). 51
Table 3.4 Linear regression results between total travel distance (complete group) and average total travel distance (CDR group). 51
Table 3.5 Correlation between radius of gyration (complete group) and average radius of gyration (CDR group). 53
Table 3.6 Linear regression results between radius of gyration (complete group) and average radius of gyration (CDR group). 53
Table 3.7 Correlation between movement entropy (complete group) and average movement entropy (CDR group). 54
Table 3.8 Linear regression results between movement entropy (complete group) and average movement entropy (CDR group). 54
Table 3.9 Summary of community detection results. 59
Table 4.1 Influence of k on the number of neighboring flows, the neighbor search runtime, the cluster merge runtime, and the number of clusters. 78
Table 4.2 Summary of population OD flows of the three tracking datasets. 83

LIST OF FIGURES

Figure 2.1 The city of Shenzhen and its six administrative districts. 18
Figure 2.2 A snapshot of STEAM visualization. It illustrates urban dynamics of Nanshan and Futian districts in Shenzhen between 7:00 and 8:00 AM. Each red point represents one hundred persons. Blue points represent cell towers. The larger the blue point, the more people are staying at its location. In this example, we select a particular cell tower near a subway station and press the "I" key. The purple lines indicate where people come from at the selected cell tower. The thickness of each purple line is proportional to the size of incoming population. 20
Figure 2.3 The stay/move dynamics in Shenzhen. 21
Figure 2.4 The cumulative distribution function (CDF) of trips captured during two selected time periods. To better illustrate short trips, this figure excludes trips greater than 50 km. 23
Figure 2.5 Temporal variation of the distance decay parameter values (β). 23
Figure 2.6 Mobility time series (stay population, incoming population, and outgoing population) of two selected cell towers. 25
Figure 2.7 Normalized average stay population of the six clusters. 26
Figure 2.8 Spatial distribution of the six clusters based on stay population. Blank cells do not cover any cell towers and most of these cells are located in areas with very limited human activities (e.g., mountains, forests). 27
Figure 2.9 Normalized average incoming population of the six clusters. 28
Figure 2.10 Spatial distribution of the six clusters based on incoming population. 28
Figure 2.11 Normalized average outgoing population of six clusters. 30
Figure 2.12 Spatial distribution of six clusters based on outgoing population. 30
Figure 3.1 Distribution of subscribers under different intensity levels of phone communication. 41
Figure 3.2 (a) Shanghai and its administrative districts. The orange areas represent the "Puxi" region, the downtown area of Shanghai. (b) Eight administrative districts in the "Puxi" region. "Puxi" and the Pudong districts are divided by the Huangpu River. 45
Figure 3.3 Temporal variation of the total number of each event. 47
Figure 3.4 Temporal variation of the total number of records in the CDR group and the complete group. 47
Figure 3.5 Distribution of subscribers under different ratio of CDRs. 48
Figure 3.6 Total travel distance (complete group) vs. average total travel distance (CDR group). 50
Figure 3.7 Radius of gyration (complete group) vs. average radius of gyration (CDR group). 52
Figure 3.8 Movement entropy (complete group) vs. average movement entropy (CDR group). 55
Figure 3.9 Cumulative distribution function (CDF) of displacements. 57

Figure 3.10 Probability density function (PDF) and the fitted power law distribution. The green line and red line represent the probability distribution of the displacements derived from the CDR group and the complete group, respectively. The dashed green line and the dashed red line are the fitted power law distributions for the CDR group and the complete group, with a decay parameter of 1.79 and 1.98, respectively. 58
Figure 3.11 Detected communities based on the CDR group. 60
Figure 3.12 Detected communities based on the complete group. 61
Figure 3.13 Decreased level of mutual exclusiveness in community detection using data from the complete group. (a) Many Voronoi cells are located outside the community boundary. (b) Taking the Voronoi cell circled in the left figure as an example, we notice that most Voronoi cells outside the community boundary cover high-density residential area. 62
Figure 4.1 10% of the OD pairs among 5,952 cell towers during 7-8 AM of a workday in Shenzhen, China. 71
Figure 4.2 An example of six OD flows of population movements. Flow 1 and Flow 2 both peak during the morning rush hours, while Flow 3 and Flow 4 have their peaks during the afternoon rush hours. Flow 5 and Flow 6, on the other hand, remain relatively consistent throughout the day. 75
Figure 4.3 Subway flow clusters using different values of k. a) k=6 (126 clusters); b) k=8 (58 clusters); and c) k=10 (41 clusters). 79
Figure 4.4 Subway flow clusters during 6-7 PM using the two-step hierarchical clustering (step 1: k=3; step 2: k=8; total number of clusters: 54). 80
Figure 4.5 a) Shenzhen and its six administrative districts; b) Locations of all subway stations; c) Locations of all cell towers in the active phone tracking data; d) Locations of all cell towers in the CDR data. 82
Figure 4.6 Hierarchical clustering results of the morning rush hour (7–8 AM) based on: a) subway smartcard data; b) active phone tracking data; c) active phone tracking data (trip distance $\geq 5km$); and d) CDR data. Note that 1) darker color represents a larger volume, and 2) to improve readability, very small OD flows which are not merged to main clusters are not displayed. 84
Figure 4.7 Flow direction distribution by administrative districts (7–8 AM). 85
Figure 4.8 Hierarchical clustering results for the evening rush hour (6–7 PM) based on: a) subway smartcard data, b) active phone tracking data, c) active phone tracking data (trip distance $\geq 5km$), and d) CDR data. 87
Figure 4.9 Flow direction distribution by administrative districts (6–7 PM). 88
Figure 5.1 The graphic user interface (GUI) of *STEAM*. Users are able to set visualization window size and data directories in the “Basic” tab (the upper figure) and visualization-related parameters in the “Visualization” tab (the lower figure). 99

Chapter 1

Introduction

1.1 Research Background and Research Questions

A city is a special form of social organization that supports various types of human activities. In the 20th century, large-scale population migration to urban areas, known as urbanization, had consolidated cities as central places for social and economic activities (Jacobs 1961; Sato and Yamamoto 2005). Human beings and cities maintain a reciprocal relationship. As pointed out by Jacobs (1961, p238), "cities have the capability of providing something for everybody, only because, and only when, they are created by everybody".

People do not move randomly in a city. Instead, human movements are largely determined by travel motivations (e.g., work/school, shopping, entertainment) and locations of urban infrastructures that can fulfill such needs. Although modern cities can provide numerous means for people to travel inside its boundary, human mobility is still restricted by factors such as distance, accessibility, schedule of operation (e.g., class time, store hours), and so forth. Hence, one has to allocate time for his/her activities conducted at different locations in the course of a day (Chapin 1974). As a result, we can, to a certain degree, feel the rhythm of human movement and observe distinct spatiotemporal patterns of urban dynamics.

A deep understanding of urban dynamics has a profound implication. On one hand, many real-world applications, such as business site selection, public transit planning, etc., require new approaches (e.g., data-driven approach) as the size of cities becomes larger. On the other hand, the rapid urbanization process has led to many urgent issues, for instance, population explosion, congestion, pollution, and so forth (Zheng et al. 2014). A prerequisite to addressing these issues is to capture the fast changing rhythm of the city. Conventional survey-based studies usually involve a very small number of participants (e.g., 1~2‰ of the population). Moreover, they are often associated with a high cost and limited spatiotemporal coverages.

In the past decade, rapidly advancing location-aware technology, information and communication technology (ICT), and the advent of the so-called "big data era" have become a game changer in urban dynamics research. A variety of new data sources have emerged and helped enhance our understanding of human mobility. They are collected in either decentralized or centralized manner. Decentralized data collection can be referred to as "crowd sourcing", or volunteered geographic information (VGI) contributed by "citizen sensors", who "create, assemble, and disseminate geographic information voluntarily" (Goodchild 2007, p211). Batty points out that such "revolution in tracking human and other motion in digital form enables the collection attributes at the finest of scale of urban observation" (Batty 2010, p575). Researchers have been using VGI, such as geo-tagged tweets and geo-tagged Flicker photos, to characterize individual or collective human mobility patterns (e.g., Preoţiuc-Pietro and Cohn 2013, Hasan et al. 2013, Azmandian et al. 2013). Centralized data collection, on the other hand, is often collected by

government agencies or companies. Typical examples include taxi GPS tracking data, public transit smartcard data, and so forth. They are usually byproducts of some real-world business operations (e.g., smartcard is designed to be an alternative way of fare collection) and carry fewer details than traditional travel surveys. Nevertheless, the large scale of digital footprints has been proved to be very useful in uncovering the pulse of a city (e.g., Pelletier et al. 2011, Liu et al. 2012, Zhou et al. 2015).

Mobile phone location data have also drawn extensive attention (Ratti et al. 2006, Candia et al. 2008). Perhaps the most prominent feature of mobile phone location data is the unprecedented scale given the pervasive use of mobile devices. Similar to many other datasets, the original intention of collecting mobile phone location data is not tracking subscribers. Instead, they are logged by mobile network operators (MNOs) for billing purposes. Every time a subscriber engages in a phone communication activity (e.g., making a phone call, receiving a text message), a call detail record (CDR) is generated to archive information such as phone numbers of caller/callee, duration of phone call, along with the ID of cell tower that handles the communication. In recent years, the value of CDRs in human mobility research has been widely recognized (e.g., González *et al.* 2008, Song *et al.* 2010a, Song *et al.* 2010b). From the individual perspective, CDRs indicate how each subscriber moves in space over time, whereas from the collective perspective, CDRs reflect how people interact with urban space at different time periods. Based on this division of analysis perspective, most existing research falls into one of the following two categories:

1) Human trajectory mining and modeling
   A large body of literature focuses on spatiotemporal characteristics of individual trajectories. Each trajectory can be described by a series of mobility indicators, such as daily range of travel, movement radius, movement entropy, etc. Based on these mobility indicators, statistical analyses can be performed to compare mobility patterns of people in different social groups (e.g., age, gender, see Kang *et al.* 2010, Yuan *et al.* 2012) or people at different locations (Becker *et al.* 2013). Individual trajectories collected over a long term allow us to extract meaningful anchor points, such as home or work locations (Ahas *et al.* 2010, Calabrese *et al.* 2010a), and to examine people's activity patterns around anchor points (Xu *et al.* 2015). Besides, CDRs provide some new insights on the nature of human travel. Over a long period of time, we believed that human movements can be explained by the random walk or the Levy flight model (Brockmann *et al.* 2006, Rhee *et al.* 2011). However, CDRs prove that human travels actually follow reproducible patterns (González *et al.* 2008) and are highly predicable (Song *et al.* 2010a, Song *et al.* 2010b).

2) Urban dynamics analysis
   Instead of focusing on individual trajectories, many studies adopt a collective approach to uncover varying mobility patterns by location. Frequently used indicators include Erlang value (i.e., the total call traffic volume in one hour), number of phone calls/text messages, number of active subscribers, etc. CDRs can be used to quantitatively measure differed levels of popularity (Girardin *et al.* 2009), or distinct patterns of mobility variation throughout different time periods in a day, or different days in a week (Reades *et al.* 2007, Calabrese *et al.* 2010b, Sagl *et al.* 2012, Yuan and Raubal 2012). Moreover, CDRs enable us to detect urban communities with strong internal interactions (Gao *et al.* 2013).

However, we should be aware of a major limitation of CDR data. That is, the collection of CDR data is event-driven (Calabrese et al. 2011). In other words, a subscriber's location is recorded only when a phone communication activity occurs. Hence, the recording frequency of different subscribers can vary significantly, depending on how actively one engages in phone communication activities. Subscribers without any phone usage are literally invisible. Without sufficient digital footprints from each subscriber, it is very challenging to examine how people stay and move at different parts of a city within a day. As a result, the time span of CDRs in existing research usually covers a relatively longer time period (e.g., a month, six months, or even longer). Although this workaround can help increase the volume of digital footprints, people who rarely use their mobile phone remain underrepresented.

Besides CDRs, some MNOs also actively track subscribers' locations by recording location of each mobile device periodically (e.g., every 30 minutes, 60 minutes, etc.), regardless of calling or texting activities. Compared with CDRs, digital footprints of most subscribers are recorded more frequently and consistently so those "silent ones" can also be traced. Furthermore, this feature allows researchers to investigate not only individual movements over time, but also time and duration people spend at each location. The latter can be valuable in many real-world applications, such as emergency evacuation, targeting of advertising campaigns, evaluation of disease transmission, etc.

Actively tracked mobile phone location data present some exciting and promising opportunities to not only uncover new insights of urban dynamics, but also re-examine and validate various findings the research community has come up with so far. This dissertation research is designed to answer the following three general research questions, each of which is further split to several specific research questions:

1) What new insights of human mobility can be gained from actively tracked mobile phone location data?
   Despite our enhanced understanding of human mobility over the past few decades, answers to many fundamental questions of human mobility are not yet clear. For instance, the spatiotemporal characteristics of move activity have been the focus of most studies, if not all of them. In reality, people do not keep moving throughout the day. Instead, for typical employees or students, daily travels only include home->work (school) commute and work (school) -> home commute during early morning and late afternoon, plus some other trips (e.g., leisure activities, shopping) along the commute trips, whereas these people probably stay at certain places (e.g., home, office/school) during the remaining time periods. What is the percentage of stay population at different times of a day? What is the relationship between stay and move activities? What are the daily variation patterns of stay activity and which urban locations share a similar pattern of stay activity? Given the limitation of social network data, taxi tracking data, and even CDRs, those questions remain unanswered. This dissertation research addresses them by taking advantages of massive hourly digital footprints that come with an actively tracked mobile phone location data. Besides stay activity, short-distance travel also plays a critical role in urban dynamics. The distance decay effects have been studied and recognized by researchers from different disciplines. However, is the frictional effect of distance always the same during different time periods of a day? If not, what is the variation pattern of distance decay effect in a city and what implication it provides for human mobility?

2) What is the bias of CDRs to reflect human mobility?
   In the big data era, there have been debates regarding the biases associated with the uneven distribution of users. For instance, studies report that distribution of social media users is predominantly uneven in terms of geography, gender, and race/ethnicity (Mislove et al. 2011, Hecht and Stephens 2014). Similarly, the representativeness of CDRs is questionable due to the uneven distribution of people's phone communication activities in space and time. On one hand, people are more likely to contact others at certain places, such as home or work location, and it is highly possible that those locations account for only a fraction of all visited places. On the other hand, depending on how actively one engages in phone communication, the total number of CDRs each subscriber generates varies significantly. Hence, we might have been overly optimistic about the usefulness of CDRs and the validness of our conclusion. How representative are CDRs in estimating individual mobility characteristics, such as total travel distance, radius of gyration, and movement entropy? Do CDRs tend to underestimate/overestimate these mobility indices? If yes, what is the level of deviation? Does such deviation vary in terms of how actively one engages in phone communication activities? In other words, do CDRs offer better

estimation for people who make a lot of phone calls, and worse estimation otherwise? In addition, what about the performance of CDRs in collective human mobility analysis, such as distance decay effect and urban community detection?

3) What are the pros and cons of different tracking datasets in urban dynamics research?
   Today, lack of data is no longer a problem in most developed regions. Instead, to answer a particular research question, we may have more than one tracking dataset and each of them reflects urban dynamics from a unique angle. In many cases, multiple tracking datasets collected in the same study area can tell different stories. As pointed out by Liu et al. (2015), such representativeness issue has become a top research priority in the big data era. It requires people to have a more thorough understanding of all available datasets in order to select the most appropriate one. This dissertation research approaches this problem by comparing population movement patterns derived from three different tracking datasets. Given tens of thousands of OD flows that overlap and clutter with each other, it is necessary to extract generalized population movement patterns. Traditionally, this is done by simply filtering out minor flows (Tobler 1987), drawing OD flows with curved lines (Wheeler 2015), or sorting flows by volume and drawing them in ascending order (Wood et al. 2011). Is there an efficient way to produce flow clusters without information loss? Is it possible to take into account the dynamics of OD flow in the clustering process? Do the three tracking datasets in this study generate similar or different population movement patterns (e.g., travel distance, travel direction)? If not, what are the strengths and shortcomings of each dataset and what are the implications to urban dynamics research?

1.2 Organization of the Dissertation

This dissertation is organized into five chapters. Chapters 2, 3, and 4 are three independent and complete manuscripts. Each of them solves one set of research questions discussed in Chapter 1.1:

Chapter 2 focuses on dynamic stay/move activities of a large city in southern China in a workday based on a large actively tracked mobile phone location dataset. An interactive visualization tool, named *STEAM* (Space-Time Environment for Analysis of Mobility), is developed to support analysis of massive digital footprints. Some important patterns revealed by *STEAM*, including the dynamic relationship of stay/move activities and the changing effect of distance decay, are then investigated. An agglomerative clustering method is applied to identify various mobility variation patterns based on three mobility indicators (volume of stay, volume of incoming population, and volume of outgoing population) and group urban areas with similar mobility patterns. The clustering analysis results further

reveal some distinct spatiotemporal patterns of urban dynamics in the selected city.

Chapter 3 investigates the bias of CDRs using a mobile phone location dataset collected from over one million subscribers in Shanghai, China. It includes CDRs (~27%) plus other cellphone-related logs (e.g., tower pings, cellular handovers) generated in a workday. All CDRs are extracted into a separate dataset in order to compare human mobility patterns derived from CDRs and from the complete dataset. From an individual perspective, the effectiveness of CDRs in estimating three frequently used mobility indicators is evaluated. From a collective perspective, both datasets are used to derive distance decay effect and urban communities. The major differences are explained by the habit of mobile phone usage in space and time. The conclusion suggests that the event-triggered nature of CDRs does introduce a considerable level of bias in human mobility patterns.

Chapter 4 proposes a revised hierarchical clustering algorithm based on an existing publication (Zhu and Guo 2014). This revised algorithm groups OD flows in terms of their proximity distance and flow similarity. Results are discussed from three aspects: 1) we summarize urban population movement patterns revealed by three tracking datasets; 2) we demonstrate how each dataset differs from others and reveals urban population movement patterns from its unique angle; 3) by combining conclusions from 1) and 2) and the characteristics of these datasets, we discuss their pros and cons in population movement analysis.

Chapter 5 provides a summary of the major contributions of this dissertation research. Possible future research directions are also outlined in this chapter.

References

Ahas R, Silm S, Järv O, Saluveer E, and Tiru M 2013 Using mobile positioning data to model locations meaningful to users of mobile phones. *Journal of Urban Technology* 17: 3–27

Azmandian M, Singh K, Gelsey B, Chang Y-H, and Maheswaran R 2013 Following human mobility using tweets. *Lecture Notes in Computer Science* 7607: 139–149

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

Brockmann D, Hufnagel L, and Geisel T 2006. The scaling law of human travel. *Nature* 439: 462–465

Calabrese F, Pereira F, Lorenzo G, Liu L, and Ratti C 2010a The geography of taste: Analyzing cell-phone mobility and social events. *Lecture Note in Computer Science* 6030: 22–37

Calabrese F, Reades J, and Ratti C 2010b Eigenplaces: Segmenting space through digital signature. *IEEE Pervasive Computing* 9: 78–84

Calabrese F, Di Lorenzo G, Liu L, and Ratti C 2011 Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Computing* 10: 36–44

Candia J, González M C, Wang P, Schoenharl T, Madey G, and Barabási A L 2008 Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical* 41: 224015

Chapin F S 1974 *Human activity patterns in the city: Things people do in time and in space*. New York: John Wiley & Sons

Gao S, Liu Y, Wang Y, and Ma X 2013 Discovering spatial interaction communities from mobile phone data. *Transactions in GIS* 17: 463–481

Girardin F, Vaccari A, Gerber A, Biderman A, and Ratti C 2009 Quantifying urban attractiveness from the distribution and density of digital footprints. *International Journal of Spatial Data Infrastructures Research* 4: 175–200

González M C, Hidalgo C A, and Barabási A-L 2008 Understanding individual human mobility patterns. *Nature* 453: 779–782

Goodchild M 2007 Citizens as sensors: The world of volunteered geography. *GeoJournal* 69: 211–221

Hasan S, Zhan X, and Ukkusuri S 2013 Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*. ACM, Chicago, Illinois, USA: 1–8

Hecht B and Stephens M 2014 A tale of cities: Urban biases in volunteered geographic information. In *Proceedings of the International Workshop on Web and Social Media*. Ann Arbor, Michigan, USA: 197–205

Jacobs J 1961 *The death and life of great American cities*. New York: Random House

Kang C, Gao S, Lin X, Xiao Y, Yuan Y, Liu Y, and Ma X 2010 Analyzing and geo-visualizing individual human mobility patterns using mobile call records. In *Proceedings of the 18th International Conference on Geoinformatics.* Beijing, China: 1–7

Liu Y, Kang C, Gao S, Xiao Y, and Tian Y 2012a Understanding intra-urban trip patterns from taxi trajectory data. *Journal of Geographical Systems* 14: 463–483

Liu Y, Liu X, Gao S, Gong L, Kang C, Zhi Y, Chi G, and Shi L 2015 Social sensing: A new approach to understand our socioeconomic environments. *Annals of the Association of American Geographers* 105: 512–530

Mislove A, Lehmann S, Ahn Y, Onnela J, and Rosenquist J 2011 Understanding the demography of Twitter users. In *Fifth International AAAI Conference on Weblogs and Social Media*

Pelletier M-P, Trépanierd M, and Morency C 2011 Smart card data use in public transit: A literature reivew. *Transportation Research Part C* 19: 557–568

Preoţiuc-Pietro D and Cohn T 2013 Mining user behaviours: A study of check-in patterns in location based social networks. In *Proceedings of the 5th Annual ACM Web Science Conference*. ACM, Paris, France: 306–315

Ratti C, Pulselli R M, Williams S, and Frenchman D 2006 Mobile landscapes: Using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design* 33: 727–748

Reades J, Calabrese, Sevtsuk A, and Ratti C 2007 Cellular census: Explorations in urban data collection. *IEEE Pervasive Computing* 6: 30–38

Rhee I, Shin M, Hong S, Lee K, and Chong S 2011 On the levy-walk nature of human mobility. *IEEE/ACM Transactions on Networking (TON)* 19: 630–643

Sagl G, Resch B, Hawelka B, and Beinat E 2012 From social sensor data to collective human behavior patterns – Analysing and visualizing spatio-temporal dynamics in urban environments. In Jekel T, Car A, Strobl J, and Griesebner G (eds) *GI-Forum 2012: Geovisualization, Society and Learning*. Wichmann Verlag, Berlin: 54–63

Sato Y and Yamamoto K 2005 Population concentration, urbanization, and demographic transition. *Journal of Urban Economics* 58: 45–61

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

Song C, Koren T, Wang P, and Barabási A L 2010b Modeling the scaling properties of human mobility. *Nature Physics* 6: 818–823

Tobler 1987 Experiments in migration mapping by computer. *American Cartographer* 14: 155–163

Wheeler A 2015 Visualization techniques for journey to crime flow data. *Cartography and Geographic Information Science* 42: 149–161

Wood J, Slingsby A, and Dykes J 2011 Visualizing the dynamics of London's bicycle-hire scheme. *Cartographica* 46: 239–251

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

Yuan Y and Raubal M 2012 Extracting dynamic urban mobility patterns from mobile phone data. *Lecture Note in Computer Science* 7478: 354–367

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

Zheng Y, Capra L, Wolfson O, and Yang H 2014 Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5: 1–55

Zhou Y, Fang Z, Thill J-C, Li Q, and Li Y 2015 Functionally critical locations in an urban transportation network: Identification and space-time analysis using taxi trajectories. *Computers, Environment, and Urban Systems* 52: 34–47

Zhu X and Guo D 2014 Mapping large spatial flow data with hierarchical clustering. *Transactions in GIS* 18: 421–435

Chapter 2

Cell Towers as Sensors: Uncovering the Pulse of a City Based on Actively Tracked Mobile Phone Location Data

Abstract

Understanding how people move or stay at different locations in a city can serve a variety of purposes, such as business intelligence and urban planning. The pervasive use of mobile phones in today's society generates digital footprints of human beings at an unprecedented scale, which allows researchers to uncover new insights of urban dynamics. Most existing research examines human mobility using call detailed records (CDRs). However, it remains challenging to analyze how people stay and move in different parts of a city within a day since the collection of CDR data is event-driven. The active tracking approach, on the other hand, records subscribers' location periodically (e.g., every hour) regardless of calling or texting activities, and thus provides a more frequent and consistent recording of individual footprints. In this paper, we study dynamic stay/move activities of a large city in southern China in a workday based on a large actively tracked mobile phone location dataset. An interactive visualization tool, named *STEAM* (Space-Time Environment for Analysis of Mobility), is developed to support analysis of massive digital footprints. Some important patterns revealed by *STEAM*, including the dynamic relationship of stay/move activities and the changing effect of distance decay, are then investigated. We then apply an agglomerative clustering method to identify various mobility variation patterns based on three mobility indicators (volume of stay, volume of incoming population, and volume of outgoing population) and group urban areas with similar mobility patterns. The clustering analysis results further reveal some distinct spatiotemporal patterns of urban dynamics in the selected city.

2.1 Introduction

A city is a special form of social organization that supports various types of human activities. Its structure influences daily population movement through “a myriad of processes” and the representation of changes in space and time is often referred to as urban dynamics (Batty 2009, p51). Urban space is not homogeneous. The socioeconomic properties (accessibility, land use, population, etc.) of each urban location largely determine its interaction with human beings (Yuan et al. 2012a). Understanding how people move or stay at different parts of a city in a day can shed light on a variety of applications, such as selecting a business site where a large number of people stay in the daytime, designing new routes for public transportation, etc.

Our understanding of urban dynamics used to be narrow due to limited data collection capabilities. Traditionally, urban planners rely on travel surveys to understand human mobility patterns (e.g., Crane and Crepeau 1998, Schlich and Axhausen 2003). However, collecting data through surveys has several drawbacks: 1) the number of participants in a survey is usually small (e.g., 1~2‰ of the population); 2) the spatiotemporal coverage of a survey is usually limited; 3) it is time-consuming and costly to collect and process survey data. From the

1990s, GPS-based tracking has been widely used to assist or replace conventional travel surveys since a GPS unit can record locations more frequently and accurately (Doherty et al. 2001, Wolf et al. 2001, Yuan et al. 2012a). Nevertheless, it is difficult to conduct extensive GPS-assisted surveys because it is costly to purchase a large number of GPS units. In the past decade, rapid developments of information and communication technology (ICT) and location-aware technology offer new opportunities for urban dynamics studies. For instance, social media apps allow people to post geo-tagged messages so human mobility in the physical space can be tracked. Batty (2010) uses Flickr and Twitter as examples to show how the pulse of a city can be studied with "crowd sourcing" data (e.g., Neuhaus 2010).

In recent years, mobile phone location data, collected by mobile network operators (MNOs), has started to draw attention in the research community. The original motivation of collecting and archiving mobile phone location data is for billing purposes (Becker et al. 2013). Every time a subscriber engages in a phone communication activity (e.g., making a phone call, receiving a text message, etc.), a call detail record (CDR) is generated to archive information such as phone numbers of caller/callee, duration of phone call, along with the ID of cell tower that handles the communication. Compared with data collected by surveys, GPS loggers, and social media apps, the scale of digital footprints collected by MNOs is unprecedented. Mobile phone location data provide some new avenues to study human-city interactions and mechanisms that drive urban dynamics (Jiang et al. 2013). Ratti et al. (2006) term this new approach of studying urban dynamics as "mobile landscapes", while Zheng et al. (2014) consider it as a critical component in the emerging realm of "urban computing".

However, we should be aware of a major limitation of CDR data. That is, the collection of CDR data is event-driven (Calabrese et al. 2011). In other words, a subscriber's location is recorded only when a phone communication activity occurs. Hence, the recording frequency of different subscribers can vary significantly, depending on how actively one engages in phone communication activities. Subscribers without any phone usage are literally invisible. Without sufficient digital footprints from each subscriber, it is very challenging to examine how people stay and move at different parts of a city within a day. As a result, the time span of CDRs in existing research usually covers a relatively longer time period (e.g., a month, six months, or even longer). Although this workaround can help increase the volume of digital footprints, people who rarely use their mobile phone remain underrepresented.

The cost of data storage today is considerably lower than it used to be. In the meantime, an increasing number of companies have realized the value of large-scale digital footprints. Therefore, besides CDRs, some MNOs also actively track subscribers' locations by recording the location of each mobile device periodically

(e.g., every 30 minutes, 60 minutes, etc.), regardless of users' calling or texting activities. Compared with CDRs, the digital footprints of most subscribers are recorded more frequently and consistently so those "silent ones" can also be analyzed. Furthermore, this feature allows researchers to investigate not only individual movements over time, but also time and duration people spend at each location. The latter can be valuable in many real-world applications, such as emergency evacuation, targeting of advertising campaigns, evaluation of disease transmission, etc.

In this study, we obtain a one-day actively tracked mobile phone location dataset collected in a major city in China. It is important to point out that individual privacy is protected because all phone numbers are anonymized. We aim to examine the changing rhythm of aggregate stay/move activities through spatiotemporal visualization and analysis. For the visualization part, a new program is desired to visualize changing stay/move activities for the following reasons: 1) some existing tools handle aggregate regional flows but lack support for the time dimension, for instance, the *FlowMap* tool (Guo 2009); 2) others deal with individual trajectories but offer limited capabilities in aggregate human mobility (e.g., *GeoTime*); and 3) the size and distribution of stay population has been ignored in existing solutions. For the analysis part, previous mobile phone data research did not focus on varying stay/move activities due to the limitations of CDRs. Empirical analysis of stay/move activities derived from actively tracked mobile phone location data can benefit our understanding of human interactions with the urban environment.

The main contributions of this paper are as follows:

- We develop a spatiotemporal visualization tool called *STEAM (Space-Time Environment for Analysis of Mobility)* for exploratory tracking data analysis, which allows users to interactively animate and query aggregated stay/move activities that change over time.
- Based on large-scale digital footprints collected every hour, we provide new insights about 1) the dynamic relationships between stay/move activities and 2) the effect of distance decay and how it varies over time in a day.
- Using an agglomerative clustering approach, we reveal different urban dynamics patterns uncovered from aggregate stay/move activities and different urban areas that share similar spatiotemporal mobility variation patterns.

The remainder of this paper is organized as follows. We review related research in Section 2.2. Section 2.3 introduces the study area and the mobile phone location data used in this study. Section 2.4 demonstrates the functions of *STEAM* and presents a video demo. Follow-up analyses regarding stay/move dynamics and distance decay dynamics are included in this section. In Section 2.5, we examine daily urban dynamics by analyzing three mobility indicators derived from massive hourly footprints. We conclude and discuss this research in Section 2.6.

2.2 Related Research

This section summarizes and discusses relevant research in the following two areas: 1) geovisualization of human mobility, and 2) mobile phone location data and urban dynamics.

2.2.1 Geovisualization of human mobility

Data visualization is a critical component in exploring and understanding a complex system. Visualizing data properly is often a crucial step for setting up meaningful research hypotheses. In the big data era, visualizing human flows and interactions, and other related variables is a challenging topic (Guo 2009, Andrienko et al. 2010). In the past, a number of visualization tools have been developed to deal with migration flows. For example, Tobler (2007) released a program called *Flow Mapper* for visualizing directions and volumes of migrations. Guo (2009) designed a method to partition all migration flows into regions and visualize regional flows. Regional flows then can be classified into groups using self-organizing map (SOM). To better handle massive movement trajectories that overlap and clutter in space, Andrienko and Andrienko (2010) developed an innovative approach to aggregate movements between small areas. Based on a large migration matrix from the 2001 UK census, Rae (2009) utilized a spatial interaction geovisualization approach to depict geographical movements associated with household mobility. Time geography (Hägerstrand 1970) also offers a powerful framework to visualize human mobility in 3D (i.e., 2-dimension of space plus the third dimension of time). Shaw et al. (2008) visualize and analyze a large individual-based migration history dataset in a time-geographic extension developed for ArcGIS.

Researchers studying urban dynamics have been making effort to design visualization approaches in order to better understand spatiotemporal movement patterns in a city. Chen et al. (2011) took a space-time GIS approach and developed an *Activity Pattern Analyst* (APA) to visualize and uncover spatiotemporal patterns from an activity-diary survey dataset. Different groups of individuals are then identified in terms of their mobility patterns over space and time. Using phone calls and GPS recordings provided by Nokia, Slingsby et al. (2013) developed interactive visualization tools to characterize spatiotemporal activity of participants' social networks. The *Real Time Rome* project illustrated the pulse of Rome by visualizing Erlang data (i.e., the total call traffic volume in one hour) in a 3-D environment (Reades 2007). The *Obama | One People* project visualized aggregate mobile phone call data in order to examine how people occupy urban space during special events, such as President Obama's Inauguration Day (Vaccari et al. 2010). The *Ville Vivante* project presented urban flows based on a large mobile phone location dataset (Schmid 2012). A prominent effect of this visualization is that population flows it generates is very similar to natural flows, such as wind (e.g., Viégas and Wattenberg 2012) or ocean circulation.

Despite these significant contributions, a new visualization tool is desired to meet the following needs of this research: 1) stay and move activities are equally important elements to be visualized; 2) support for the time dimension so the changing rhythm of the city can be examined; and 3) capabilities to visualiza and query aggregate stay/move activities. These requirements are our motivations to develop *STEAM*, an interactive visualization tool for exploratory analysis of tracking data.

2.2.2 Mobile phone location data and urban dynamics

The emerging field of mobile phone data analytics has received significant attention and its value has been widely recognized. Existing literatures are mostly based on CDRs which, in general, fall into two categories in terms of the perspective of analysis: the individual/group perspective and the urban system perspective.

Many studies approach urban dynamics from an individual/group perspective, which considers the trajectory of each subscriber as a basic unit. Using conventional statistical methods, the large volume of individual trajectories has led to many interesting findings, such as similar/dissimilar mobility patterns among different population groups (e.g., age, gender, etc., Kang et al. 2010, Yuan et al. 2012b). It also provides opportunities to explore basic rules that govern urban dynamics, such as activity space (Yuan et al. 2012b, Kang 2012a), distance decay effect (Kang et al. 2012a, Kang et al. 2013, Gao et al. 2013a), predictability of human mobility patterns (González et al. 2008, Song et al. 2010a), scaling properties of human mobility (Song et al. 2010b), etc. Although to what extent human mobility can be predicted remains controversial, we are capable of identifying the frequently visited locations (e.g., home, work) for the majority of users with a high level of confidence (Calabrese et al. 2010a, Berlingerio et al. 2013, Xu et al. 2015). On the application side, knowledge we gain about individual travel routines can be applied to detect abnormal mobility patterns and massive social events (Traag et al. 2011, Ferrari, et al. 2014).

Instead of taking an individual/group perspective, some researchers pay more attention on the urban system and examine the varying intensity of human activity across urban space. With this approach, a study area is usually divided into subregions, such as grid cells, or Voronoi polygons. Physical movements or virtual communications in each subregion can be aggregated to derive indicators of interests, such as frequency of phone calls (Yuan and Raubal 2012), Erlang values (Reades et al. 2009, Gao et al. 2013b), and number of active subscribers (Kang et al. 2012b). Some studies went further to probe the correlation between human mobility and socioeconomic environment, such as land use types (Calabrese et al. 2010b, Pei et al. 2014) or social events (Calabrese et al. 2010a).

These studies have deepened our understanding of human mobility and urban dynamics based on a data-driven approach. However, most findings are within the scope of people's communication activities. Some fundamental aspects, such as how people stay, come to, and leave each urban location over time has not been carefully examined. In this research, we adopt a new approach and derive several mobility indicators from actively tracked mobile phone location data, which measure the changing urban dynamics at different urban locations. These valuable indicators form the basis of our empirical analysis.

2.3 Dataset

The mobile phone location data used in this study is obtained through a joint research collaboration. This section introduces the study area and the dataset.

2.3.1 Study area

Our study area is Shenzhen, a large city in southern China. Shenzhen has a permanent resident population of more than 10 million by the end of 2012 (see Gazette of the People's Government of Shenzhen Municipality 2012). In addition to permanent residents, Shenzhen has a large population of migrants, as it is one of the centers of manufacturing industry in China. Shenzhen was designated as a Special Economic Zone (SEZ) in 1980 and its economy has been growing rapidly since then. As of 2012, Shenzhen's annual GDP ranked the 4th in China (after Beijing, Shanghai, and Guangzhou, see Gazette of the People's Government of Shenzhen Municipality 2012). Shenzhen consists of six administrative districts: Luohu, Futian, Nanshan, Baoan, Longgang, and Yantian (Figure 2.1). Luohu is connected with Hong Kong and it is the financial center of Shenzhen. Futian is at the center of the Shenzhen Special Economic Zone. It is also considered as the business center of Shenzhen. Nanshan is a district designated for higher education and advanced technology. Baoan occupies a large area in western Shenzhen. It is an ideal region for many manufacturing industries. The largest factory of Foxconn, manufacturer of Apple's iPhone, is located in this district. Longgang is the largest district in Shenzhen. It is also a preferred region for manufacturing industries and it meanwhile is the base for export and logistics industry. On the east side of the Shenzhen Special Economic Zone is Yantian, where a major port of Shenzhen and many costal resorts are located.

As a young and emerging major city in China, the government of Shenzhen collects and utilizes various types of tracking data for urban planning, including taxi tracking data, smartcard data, and so forth. Some interesting aspects of Shenzhen urban dynamics have been reported in the literature. For instance, Liu et al. (2009) analyzed a smartcard dataset and indicated the dominant role of certain subway stations of Shenzhen in terms of passenger volumes. Zhu and Guo (2014) developed a hierarchical clustering algorithm to analyze taxi flows and suggested distinct urban dynamics patterns in the morning and evening hours, respectively.

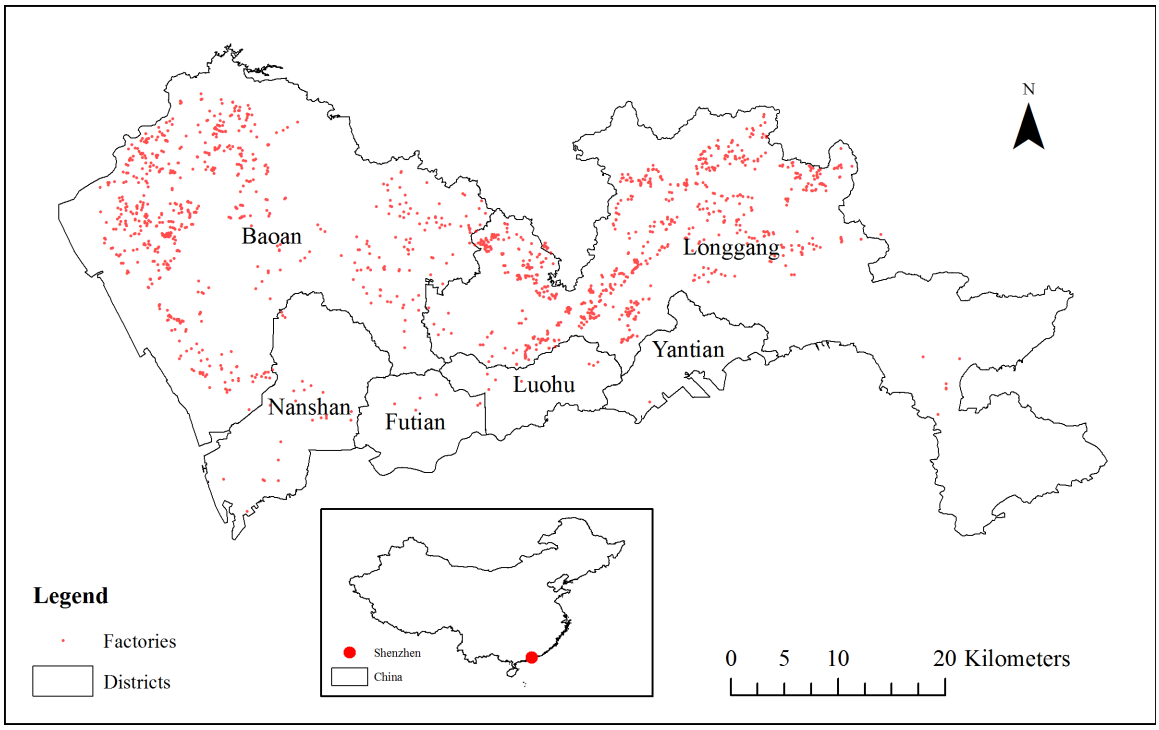

Figure 2.1 The city of Shenzhen and its six administrative districts.

2.3.2 Dataset

This actively tracked mobile phone location dataset contains locational information of over 16 million anonymized phone numbers between 23:00 of a Thursday and 22:00 of a Friday in 2012. Different from CDRs, location of most subscribers in this dataset is recorded every 60 minutes or so, as the latitude and longitude of a nearby cell tower, regardless of their calling or texting activities. From those latitudes and longitudes, we extracted 5,952 cell towers in Shenzhen and assigned each of them a unique ID.

To study Shenzhen's urban dynamics on this particular workday, we grouped digital footprints into 23 one-hour time periods. For each one-hour time period, we aggregated subscribers and derived two aspects of urban dynamics: *stay* activity and *move* activity. Specifically, we compute the size of stay population at each cell tower, and the size of bi-directional flows between each pair of cell towers.

Although this dataset provides hourly digital footprints from millions of subscribers, some data uncertainty issues should be noted. In general, uncertainties are often caused by the following two factors:

1) Distribution of cell towers: The spatial granularity of this mobile phone location dataset is at the cell tower level. Therefore, very short local trips

are not captured in the dataset. In addition, the density of cell towers varies across the city. Such short movements are less likely to be "sensed" in regions where cell towers are sparsely distributed.

2) Sampling rate: The one-hour sampling interval of this dataset implies that round trips with a very short time duration are not captured. As a result, a subscriber who leaves the service area of a cell tower and returns to the same service area within a one-hour time window is treated as *stay* during that hour.

We should keep these data uncertainty issues in mind as they cannot be eliminated due to the characteristics of this dataset. At the urban scale, we believe such uncertainties have a limited impact on the aggregate urban dynamics patterns.

2.4 Examine urban dynamics with *STEAM*

To support this empirical study, we developed *STEAM* to visualize aggregate stay/move activities based on tens of millions of digital footprints. *STEAM* was developed with *Processing*, an open source programming language based on Java and an integrated development environment (IDE) for art and visual design (Reas and Fry 2010). In addition to the basic graphic functions, *Processing* is well known for its powerful support for animation and user interaction. We also used a third-party library, *MapThing*, to embed geovisualization in *Processing* (Reades 2013). The main functions of *STEAM* are summarized as follows:

- *STEAM* represents population movement by drawing moving points between origins and destinations (Figure 2.2). The number of moving points between an OD pair is directly proportional to the size of population moving between two locations. Users can specify the number of people each moving point represents.
- *STEAM* represents population stay by drawing fixed points (Figure 2.2). The size of each fixed point is directly proportional to the magnitude of population staying at a particular location. Users can specify the number of people each point size represents.
- With stay/move activities for multiple time periods, users can move forward and backward along the time dimension.
- *STEAM* supports interactive queries. By hovering mouse cursor over a cell tower and pressing selected keys, *STEAM* shows the sizes of staying, incoming, and outgoing population at the selected cell tower. It also draws lines to indicate the origins of incoming population and the destinations of outgoing population (Figure 2.2).

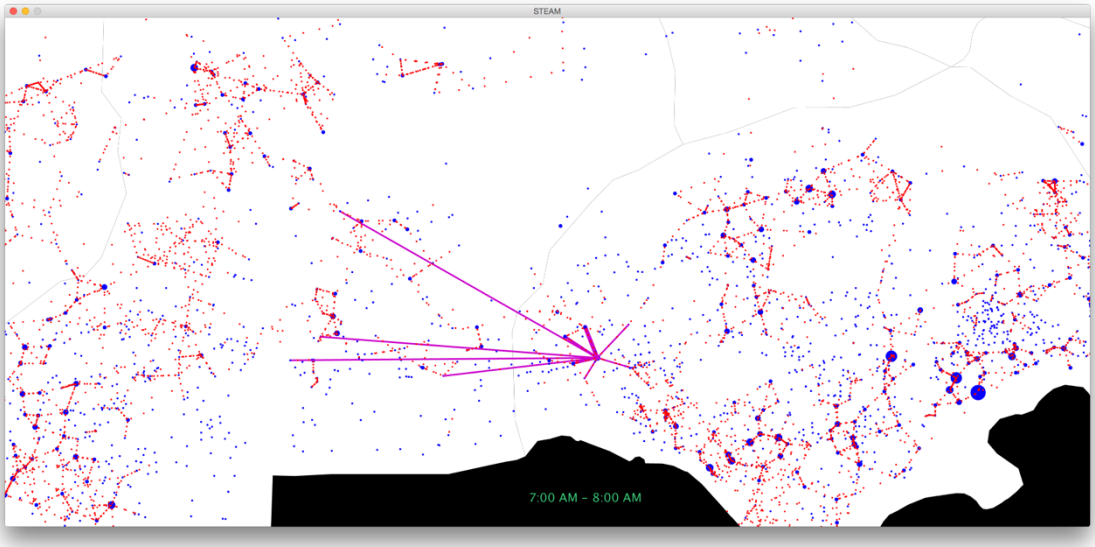

Figure 2.2 A snapshot of STEAM visualization. It illustrates urban dynamics of Nanshan and Futian districts in Shenzhen between 7:00 and 8:00 AM. Each red point represents one hundred persons. Blue points represent cell towers. The larger the blue point, the more people are staying at its location. In this example, we select a particular cell tower near a subway station and press the "I" key. The purple lines indicate where people come from at the selected cell tower. The thickness of each purple line is proportional to the size of incoming population.

2.4.1 Stay/move dynamics

STEAM reveals distinct mobility patterns throughout the day based on the varying number of moving points (i.e., move) and the varying size of fixed points (i.e., stay). This actively tracked mobile phone location data can help quantify this changing relationship between stay and move activities. We calculated the total population who stayed at the same cell tower, and those moved from one cell tower to another (regardless of distance), for each of the one-hour time intervals (Figure 2.3). In general, we notice that stay population is always more than move population throughout the study day. It highlights the importance of stay activity in an urban area, which often does not receive sufficient attention in urban dynamics studies.

Starting from 1:00 AM, the proportion of move population is less than 15% since most people stay at home during this time period. This number drops continuously when people return home gradually. The mobility level is elevated significantly from 6:00 AM. A considerable amount of people commute to work during the morning rush hour (7:00 – 8:00) as we find that over 46% population travel to the service area of a different cell tower. The proportion of stay population is relatively consistent from 8:00 to 17:00, ranging from 57.21% to 63.89%. It indicates a typical 8:00 to 17:00 workday schedule. The small growth of move population

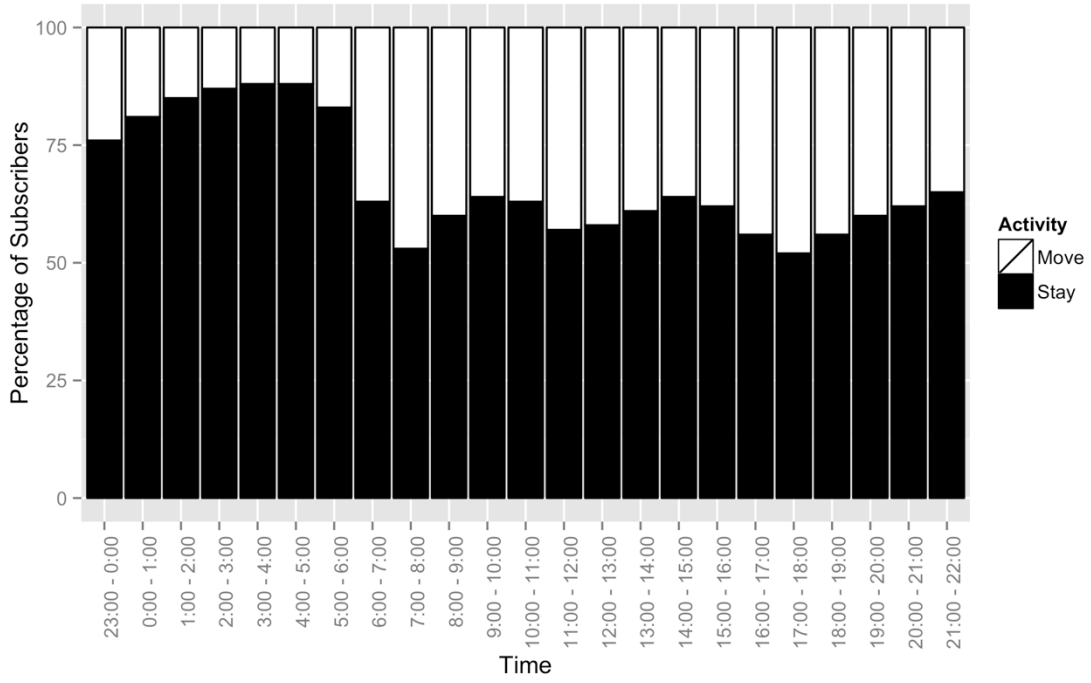

Figure 2.3 The stay/move dynamics in Shenzhen.

(42.79%) at 11:00 AM – 12:00 Noon suggests a mild urban mobility increase around noon and it echoes what we observe in *STEAM*. Unlike many other countries, a two-hour lunch break is a common practice in China. We believe the observed mobility increase at noon is the result of people going out for lunch or returning home during the lunch break. The proportion of move population increases and reaches the daily maximum (48.11%) during 17:00 – 18:00. The mobility level then declines steadily.

2.4.2 Distance decay dynamics

Despite an increased volume of long-distance travels during rush hours, *STEAM* discloses the dominant role that short-range flows play in Shenzhen's urban dynamics. It reveals a strong distance decay effect. Previous research based on CDRs has suggested various effects of distance decay in urban areas. For instance, González et al. (2008) and Gao et al. (2013a) report distance decay parameters of 1.75 and 1.60, respectively. As opposed to CDRs, the actively tracked mobile phone location data allows us to examine the varying distance decay effect throughout the day based on massive digital footprints collected from all subscribers, even if they do not engage in any phone communication activities.

We aggregated all trips that occurred during each hour and represented trip distances by a power law distribution as follows:

$$P(d) \propto d^{\beta}$$

where β is the distance decay parameter. A small value of β indicates a relatively weak influence of distance, while a large value of β implies that distance is a strong deterrent to spatial interaction.

Figure 2.4 shows the cumulative distribution function (CDF) of trips captured during two selected time periods (3:00 – 4:00 and 15:00 – 16:00) when β has the smallest value and the largest value. Both curved lines indicate that most trips are short: around 87% trips at 15:00 – 16:00 are shorter than 5 km and this percentage is even higher (> 95%) for trips during 3:00 – 4:00. The CDF value grows more quickly at 3:00 – 4:00 as distance increases, and the fitted power law has a larger value of β (2.369). This indicates that the friction of distance has a stronger effect for people traveling at 3:00 – 4:00. On the contrary, during 15:00 and 16:00, a higher percentage of people travels over a longer distance and it leads to a smaller value of β, which is very close to 1.75, as reported by González et al. (2008).

The plot of varying distance decay parameter values (Figure 2.5) indicates that the distance decay effect is relatively weak in the daytime. We notice that the decay parameter rises at noon. This finding, again, matches what we uncover in *STEAM* and the finding in Section 2.4.1, which indicates that more moving points emerge at noon but most of them are between neighboring towers. It is also interesting to see that the minimum β occurs for trips at 15:00 – 16:00, instead of typical rush hours when the demand for longer distance travel is supposed to be higher. This could be due to traffic congestion during rush
hours that restricts the distance people can travel within an hour. Finally, we find that the value of β on average is much smaller than those reported by CDR-based studies, especially during 23 PM and 5 AM. It serves as an indication that short-range movements are underrepresented in CDR data. As a result, CDR data may underestimate the urban distance decay effect and derive misleading urban mobility patterns.

2.5 Uncover urban dynamics using mobility time series

In Section 2.4, we explored some interesting aspects human mobility pattern at the urban scale. As indicated by *STEAM*, actively tracked mobile phone location data allow us to examine varying mobility patterns in different areas of the city at a much finer spatiotemporal granularity. This section presents urban dynamics patterns identified from analysis of mobility time series data.

2.5.1 Method

We selected two cell towers from a residential area and the CBD area, respectively, and computed three mobility time series: stay population, incoming

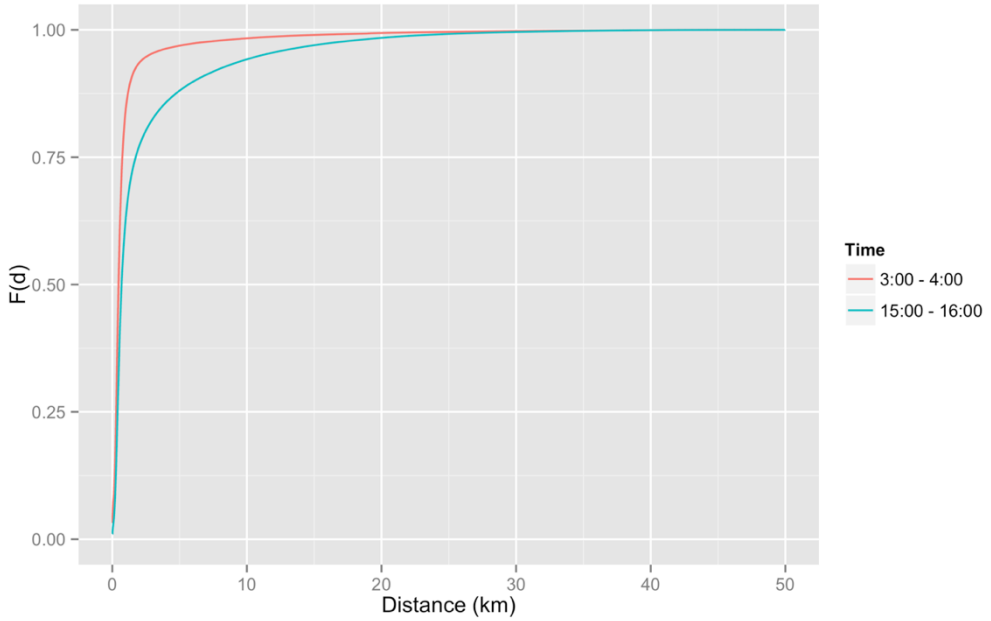

Figure 2.4 The cumulative distribution function (CDF) of trips captured during two selected time periods. To better illustrate short trips, this figure excludes trips greater than 50 km.

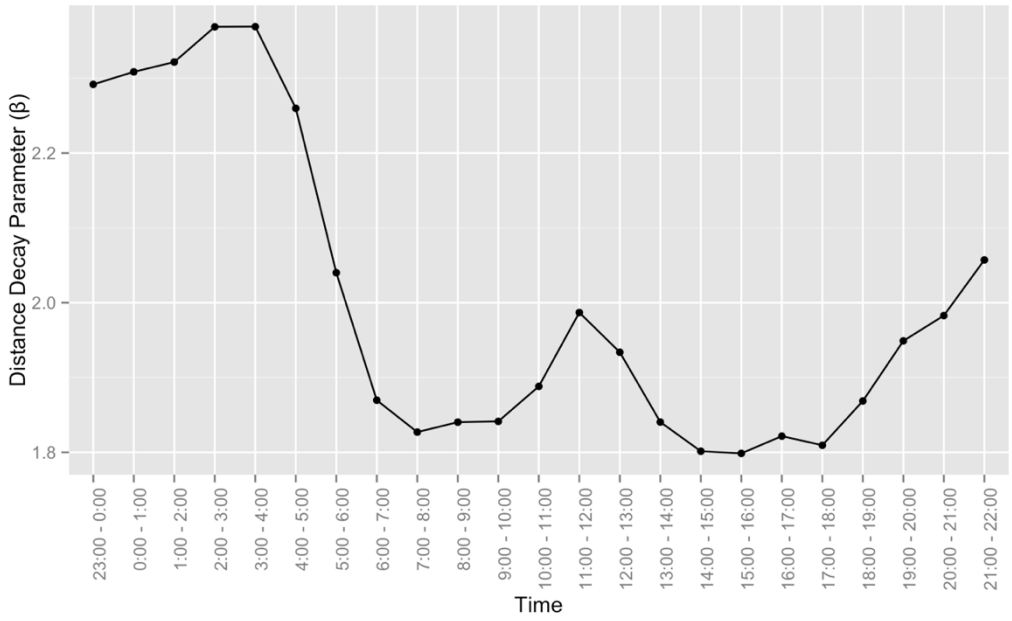

Figure 2.5 Temporal variation of the distance decay parameter values (β).

population, and outgoing population. In order to study the variation pattern of mobility, we normalized the time series data so each of them has a mean of zero and a standard deviation of one (Lin et al. 2003). The plot of these mobility time series suggests distinct mobility variation patterns of two typical types of urban location (Figure 2.6):

1) The level of *stay* population at the selected tower in the residential area is high at midnight and starts to drop during 6:00 – 7:00. It stays at a low level until 17:00 – 18:00 when people start to return home. On the contrary, the stay population at the selected CBD area shows the opposite variation pattern, which has a high level during the work hours and the number of people staying in this region declines from 16:00 – 17:00.
2) The size of *incoming* population at the selected tower in the CBD area peaks during 8:00 – 9:00. After that, fewer and fewer people come to this area (except a mild increase around noon). The cell tower in the residential area has three spikes of population gains at 11:00 – 12:00, 17:00 – 18:00, and 21:00 – 22:00, which correspond to returning population during the noon break, returning population after work, and returning population after night activities, respectively.
3) The level of *outgoing* population at the selected tower in the CBD area reveals a typical workday pattern. It shows limited growth at noon and peaks during the evening rush hour when most people get off work. Two major spikes of the same indicator happen in the selected residential area during two time periods when many people leave for work in the morning and after the noon break, respectively.

The mobility time series analysis discussed above is used to uncover aggregate urban dynamics patterns for the entire city of Shenzhen. We first subdivided Shenzhen into a grid of 2,193 cells (1 km $\times$ 1 km each) to overcome the uneven spatial distribution of cell towers. The 5,952 cell towers, as well as the hourly stay/incoming/outgoing population at each tower, were then assigned to corresponding grid cells. Three normalized mobility time series (i.e., stay/incoming/outgoing population) were derived for each cell. To identify different variation patterns of urban dynamics and examine which cells share a similar daily mobility variation pattern, we applied an agglomerative clustering approach that has been frequently used to classify time series data (Liao 2005, Tan et al. 2005).

2.5.2 Results

This section presents the clustering results of three different mobility time series. We chose six as the number of clusters after many trial runs. When the number of clusters is smaller than six, it is insufficient to distinguish all major urban dynamics patterns. When the number of clusters is larger than six, we begin to have clusters that are too similar to each other.

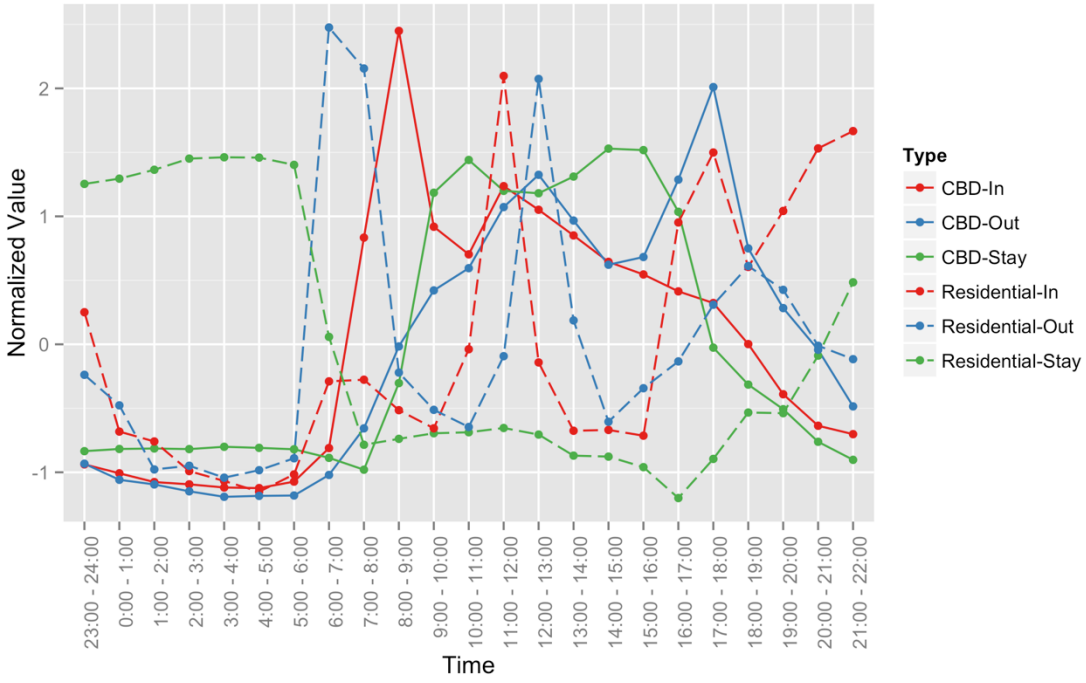

Figure 2.6 Mobility time series (stay population, incoming population, and outgoing population) of two selected cell towers.

2.5.2.1 Stay population

Understanding the *stay* population distribution in different time periods of a day has a profound meaning in the real world. For instance, it can benefit the local government in emergency evacuation planning. On the business side, restaurants and other services may be interested in areas with a large stay population during the daytime to attract customers.

The six clusters derived from the agglomerative clustering method suggest the following urban dynamics patterns of stay population (Figure 2.7). The stay population variation of both Cluster #5 and #6 reveals a typical pattern of night-stay locations. The level of stay population starts to drop from 5:00 – 6:00 and then declines dramatically during the morning rush hours until 7:00 – 8:00. Starting from 18:00 – 19:00, stay population of these two clusters increases at a nearly constant rate, indicating that people return home steadily. A major difference between these two clusters is that, throughout the daytime, stay population of Cluster #5 stays low while there are two small growth periods for Cluster #6. By overlaying the cluster map on the aerial photo in *Google Earth*, we find that high-density residential buildings are more dominant in cells in Cluster #5 than those in Cluster #6. In other words, grid cells in Cluster #6 tend to covers other types of facilities where people stay during work hours, such as shopping malls, office

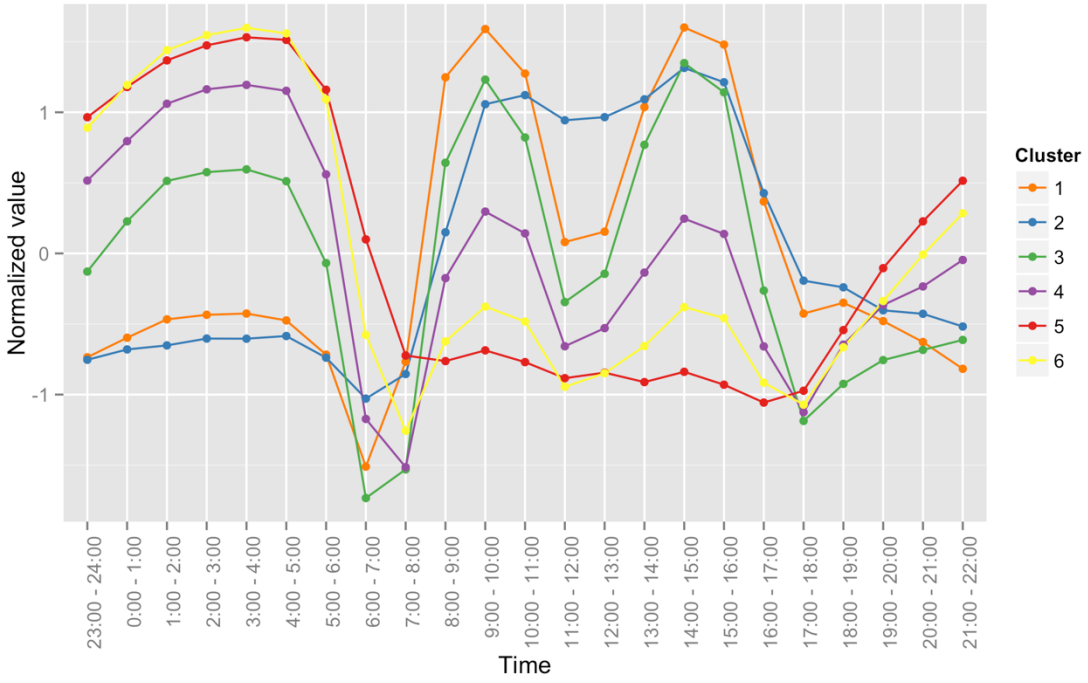

Figure 2.7 Normalized average stay population of the six clusters.

buildings, or factory buildings. Figure 2.8 shows that these night-stay locations are widely scattered in the city. Grid cells in Cluster #5 are more concentrated in the south of the city, where residential areas with tall buildings are more densely distributed.

Clusters #1 and #2 can both be regarded as day-stay locations since their level of stay population is high only during work hours (Figure 2.7). However, Cluster #1 differs from Cluster #2 in that its stay population drops significantly during the noon break. On the contrary, only a slight drop of stay population is found in Cluster #2 cells. It implies two main types of activities at noon: staying at the work place, and traveling to another location (probably going for lunch or returning home). The cluster map (Figure 2.8) indicates that the former are more common among people working in the southern part of Shenzhen, where high-tech and other tertiary industries are concentrated, such as universities and office buildings for banks, finance services and IT sectors.

Clusters #3 and #4, which scatter around the city without an evident agglomeration (Figure 2.8), present another stay population variation pattern (Figure 2.7). The level of stay population in these clusters is low during rush hours and a major drop occurs at noon. It fits the overall variation pattern of stay population in the entire city of Shenzhen (Figure 2.3). Cross-comparison with aerial photos offers two possible explanations. First, very mixed land use patterns (e.g., residential area

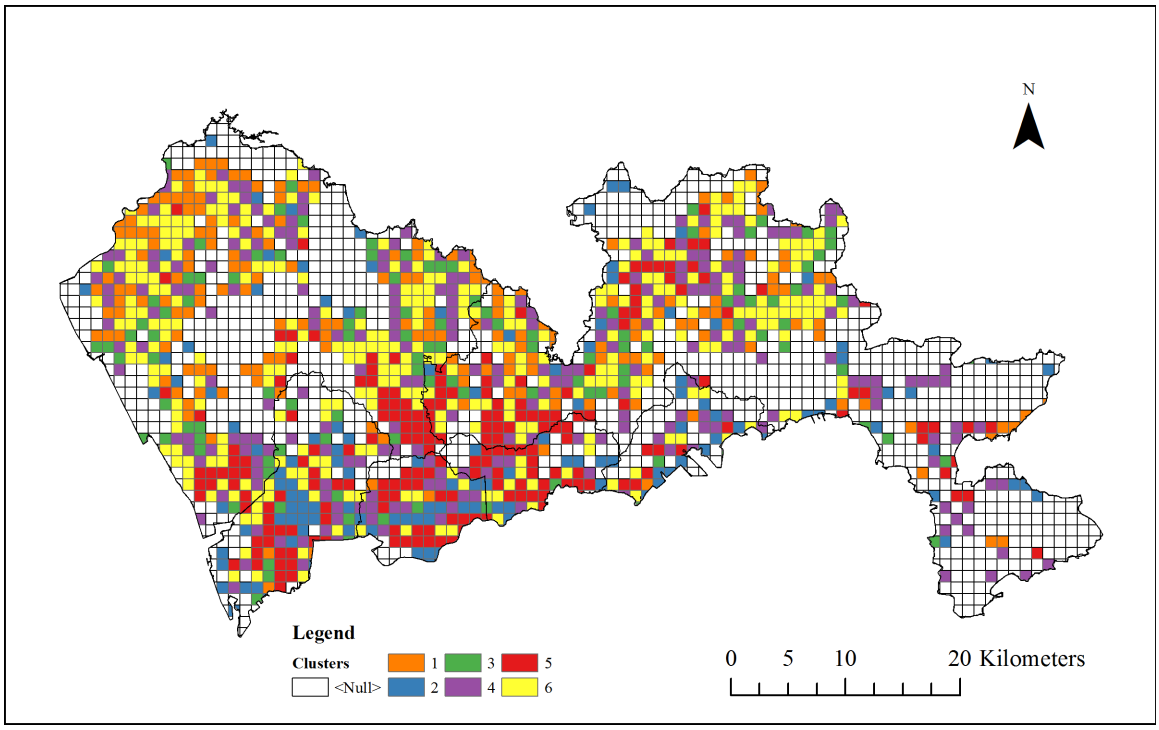

Figure 2.8 Spatial distribution of the six clusters based on stay population. Blank cells do not cover any cell towers and most of these cells are located in areas with very limited human activities (e.g., mountains, forests).

and work area) are found in grid cells of these two clusters. As a result, the level of stay population is low during the rush hours when the majority of people are travelling. Second, some grid cells in these two clusters cover the least populated areas, such as farmlands. Our speculation is that the variation of stay population in these regions is largely influenced by people who live and work there.

2.5.2.2 Incoming population

The size of incoming population during each hour indicates the speed of population gain and it thus reflects the changing attraction of a region. An enhanced understanding about how fast each urban location gains population throughout a day can assist urban planning such as optimizing bus/subway schedules.

Among the six clusters, Clusters #1 and #6 share a similar incoming population variation pattern (Figure 2.9). Apparently the three spikes of incoming population can be considered as the main characteristic of these two clusters, although locations in Cluster #1 have a greater population gain during the morning rush hour. Most grid cells of these two clusters are distributed in the northern part of the city, where is a popular region for manufacturing industries (see Figure 2.10 for factory locations in Shenzhen). The reasons of this three-spike pattern are two-fold. First, spatial adjacency of factory buildings and residential buildings is very

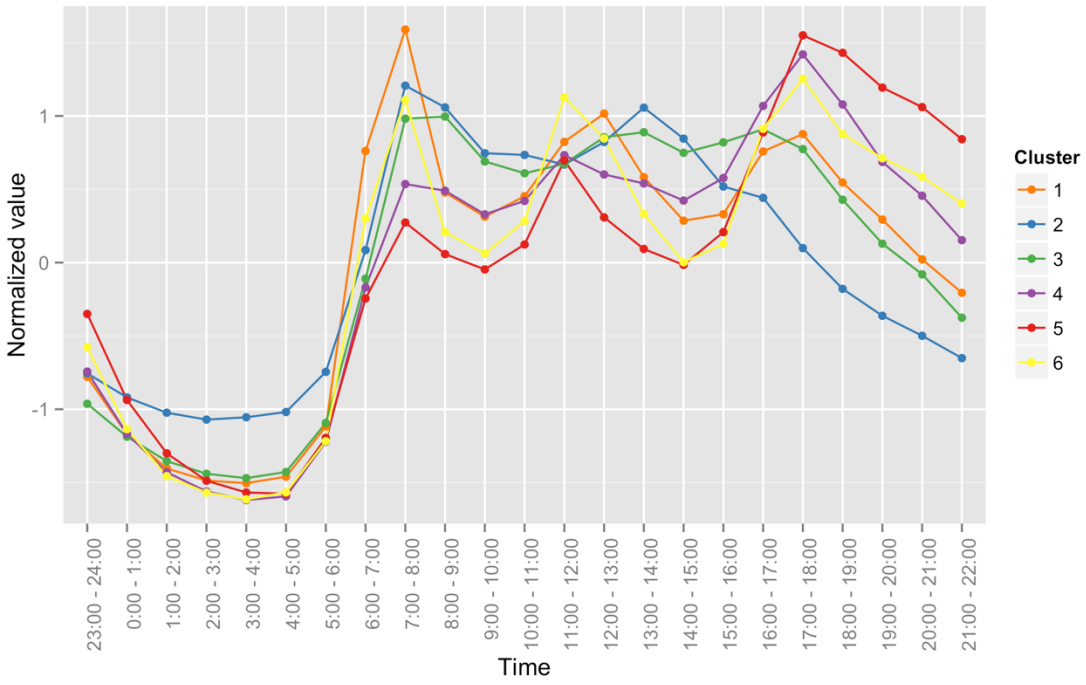

Figure 2.9 Normalized average incoming population of the six clusters.

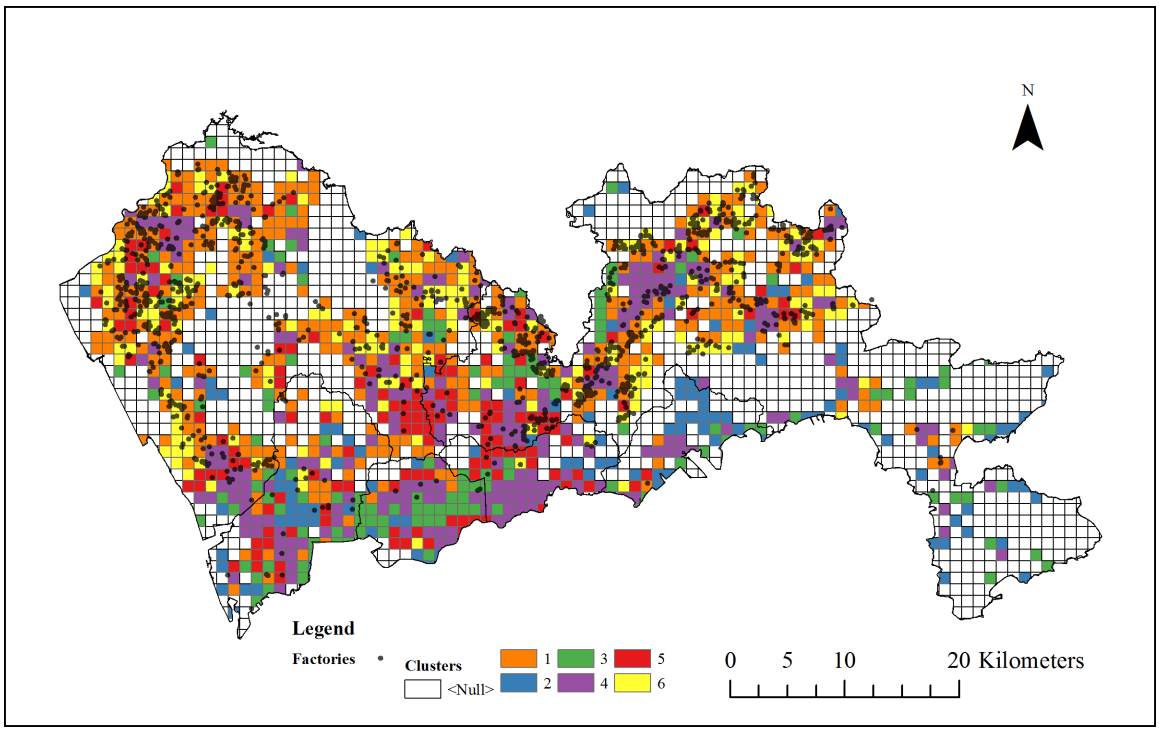

Figure 2.10 Spatial distribution of the six clusters based on incoming population.

common in this area. The three-spike pattern can be explained by a large number of people in northern Shenzhen commuting to nearby workplaces in the morning and returning home (possibly dormitories provided by factory) at noon break and in the evening. Second, we speculate that, since many manufacturing factories in China often operate beyond normal work hours and many factories have three shifts in a day, workers may commute to the factories in the evening to work on a night shift. This may also contribute to the elevated level of incoming population in the evening.

Clusters #4 and #5 also gain population at noon and during the rush hours, while the major increase of both clusters occurs during the evening rush hour (Figure 2.9). Most grid cells of these two clusters (especially Cluster #5) overlap with areas of high-density residential buildings, which attract a large population after work hours.

The agglomerative clustering results also indicate the distinct variation pattern of incoming population in the more economically developed regions of Shenzhen. The business-oriented areas in Futian and west Luohu are largely covered by the grid cells of Cluster #3 (Figure 2.10). The level of incoming population of this cluster is high, but stays quite consistently, during the daytime (Figure 2.9). This consistency implies that no significant increase of movement to these regions. It is likely that a large fraction of people working in this region do not make long-distance trips (> 2 km) at noon, which further confirms our findings in Section 2.5.2.1 that the level of stay population of Cluster #2 only drops slightly at noon. On the other hand, Cluster #2 shows some limited variation of incoming population after the noon break (Figure 2.9). Grid cells of Cluster #2 cover the central area of Nanshan, where Shenzhen University and other high-tech campuses are located, along with many tourist resorts (e.g., mountains, beaches) in the eastern part of Shenzhen.

2.5.2.3 Outgoing population

Different from incoming population, outgoing population in each hour measures the speed of population loss. This mobility indicator can also be useful for many real-world scenarios. For instance, taxi companies may benefit by dispatching drivers to areas with a substantial outgoing population.

Compared with the previous two indicators, the six clusters based on this mobility time series (Figure 2.11) present a stronger effect of spatial agglomeration: grid cells of the same or similar clusters are more likely to be adjacent (Figure 2.12). For instance, Cluster #2 covers the three more developed districts in the south. The major increase of outgoing population in these areas is probably due to people who get off work in the evening (Figure 2.11). In addition, the level of outgoing population of Cluster #2 has a very limited growth at noon, which is another proof that long distance travel during the noon break is not a common practice for people

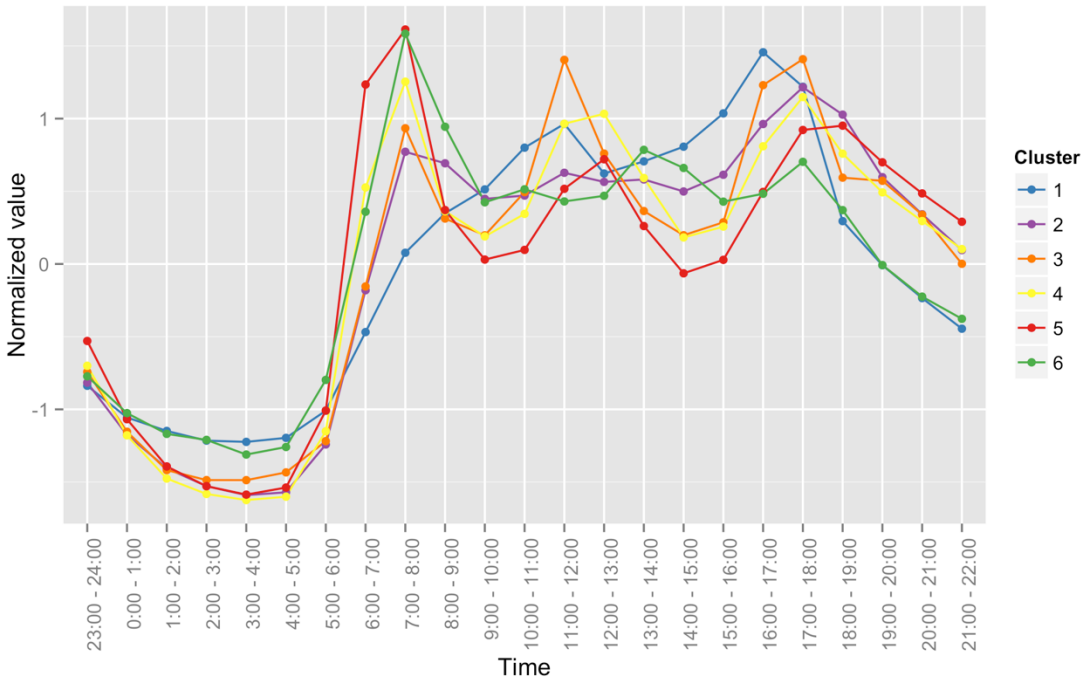

Figure 2.11 Normalized average outgoing population of six clusters.

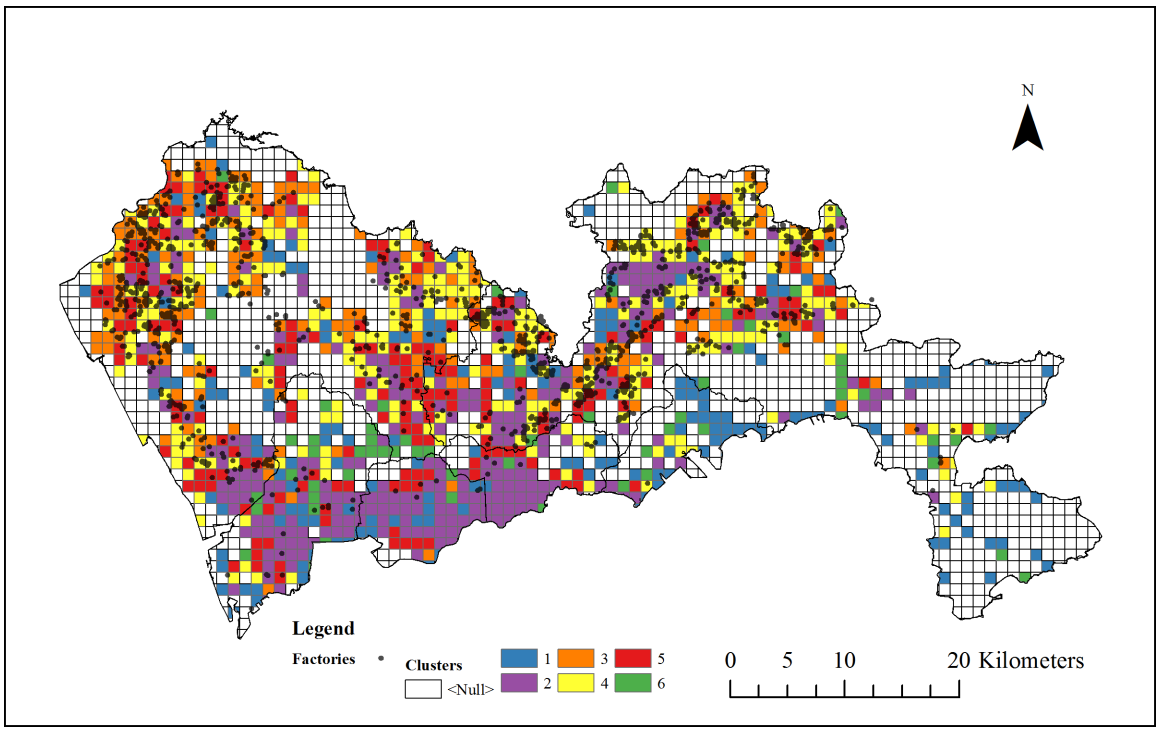

Figure 2.12 Spatial distribution of six clusters based on outgoing population.

working in this region.

Despite some slight differences, Clusters #3, #4, and #5 all share a three-spike pattern of outgoing population (Figure 2.11). Grid cells of these three clusters are more spatially agglomerated in northern Shenzhen, where manufacturing factories are densely located (Figure 2.12). As opposed to Cluster #2, a lot more people leave grid cells of Cluster #3, #4, and #5 at noon for some purpose (e.g., retuning home). Such distinct noon activity patterns between the south and the north is noteworthy.

Cluster #1 is different from others because of its low level of outgoing population during the morning rush hour (Figure 2.11). A cross-comparison with aerial photo indicates that grid cells of this cluster barely overlap with residential areas. On the contrary, most cells in Cluster #6 cover high-density residential buildings, resulting in a striking growth of outgoing population in the morning rush hour.

Actively tracked mobile phone location data allows urban dynamics patterns to be examined with improved spatiotemporal granularity. In this study, the volume of total phone number (16 million) is at the same scale as the entire population and the consistent one-hour sampling rate helps reveal aggregate stay/move activities throughout the day. Taking advantages from those massive hourly footprints, the derived mobility time series (i.e., stay population, incoming population, and outgoing population) provides good estimates of the changing pulse of the city. In addition, the agglomerative clustering approach is proven to be effective in extracting distinct mobility variation patterns from mobility time series and grouping areas with similar patterns. Provided with sufficient digital footprints, this proposed approach can be applied to reveal useful urban dynamics patterns in other cities.

2.6 Conclusions

In recent years, CDRs have been a useful data source to study urban dynamics. Unlike CDRs, actively tracked mobile phone location data collected by MNOs provide more consistent records of stay/move activities in space and over time, regardless of their calling or texting activities. We argue that cell towers work as sensors, which form an enormous sensor network and keep monitoring the changing pulse of a city around the clock. The enhanced spatiotemporal granularity of actively tracked data can help us gain additional insight regarding the ways people interact with different urban areas.

Based on the hourly digital footprints collected from over 16 million phone numbers, this study used an aggregate perspective to uncover hidden urban dynamics in Shenzhen. A new visualization tool, *STEAM*, was developed to illustrate the pulse of the city. By aggregating stay/incoming/outgoing population at each cell tower, *STEAM* uncovered some interesting characteristics of urban

dynamics in a workday in Shenzhen, such as the variation of stay/move activities and the dominant role of short-distance movements. Assisted by *STEAM*, we further performed quantitative analyses to investigate the dynamic relationship of stay/move activities and the changing effect of distance decay. *STEAM* can support other types of tracking data, such as GPS tracking data, public transit smart card data, to name a few, and it is released as an open source tool (https://github.com/zlzhao1104/steam).

To identify the variation patterns of urban dynamics in different areas of the city in a workday, we divided Shenzhen into grid cells and selected three mobility indicators (i.e., stay population, incoming population, and outgoing population) to measure urban dynamics from three different aspects. Empirical results generated by agglomerative clustering presented some interesting findings of Shenzhen's urban dynamics, such as the distribution of day-stay locations and night-stay locations, different mobility patterns between the south and the north as well as between residential areas and CBDs. We believe that aggregate stay/move mobility patterns are closely related to the urban structure, especially the setting of land use or industrial types in each region. This is particularly true in Shenzhen, where distinct industrial distributions are present in different administrative districts. Compared with the conventional data and methods, this proposed approach can be useful for urban planners and policy makers to understand varying mobility patterns at a high spatiotemporal granularity.

By discussing empirical results from spatiotemporal visualization and analysis, this paper demonstrates the usefulness of actively tracked mobile phone location data in urban dynamics study. Nonetheless, certain limitations of this study should be noted. First, a detected movement from one's trajectory does not necessarily represent the true origin and the true destination of a trip. Instead, it could be an in-transit point during a trip. Thus, the derived number of incoming/outgoing population of a location may include people passing through a location. Second, the three mobility indicators are investigated independently. A future study can examine the relationships among the three mobility indicators at each location. Third, population flows are aggregated at tower-to-tower level. Road networks and travel modes are not considered in this study, which limits its usefulness of addressing specific transportation planning tasks.

There remain some interesting and challenging topics for future research. For instance, various types of tracking data have been collected in cities like Shenzhen (e.g., taxi tracking data, public transit smartcard data, and mobile phone tracking data) and they reflect different aspects of urban dynamics. Evaluating the strengths and weaknesses of each type of tracking data can help urban planners choose the most appropriate dataset to address a specific question. Also, it will be promising to develop additional indicators to summarize the characteristics of each subscriber (e.g., staying at one location all day, traveling a lot during the

daytime/midnight, etc.), or examine the relationships between mobility patterns and the characteristics of different urban location such as accessibility, land use, average annual household income, etc. This will further improve our understanding of different mobility patterns and the intrinsic mechanisms that drive urban dynamics.

References

Andrienko G, Andrienko N, Demsar U, Dransch D, Dykes J, Fabrikant S, Jern M, Kraak M J, Schumann H, and Tominski C 2010 Space, time, and visual analytics. *International Journal of Geographical Information Science* 24: 1577–1600

Andrienko G and Andrienko N 2010 A general framework for using aggregation in visual exploration of movement data. *The Cartographic Journal* 47: 22–40

Batty M 2009 Urban modeling. In Thrift N and Kitchin R (eds) *International Encyclopedia of Human Geography*. Oxford, UK: 51–58

Batty M 2010 The pulse of the city. *Environment and Planning B: Planning and Design* 37: 575–577

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

Berlingerio M, Calabrese F, Di Lorenzo G, Nair R, Pinelli F, and Sbodio M L 2013 AllAboard: A system for exploring urban mobility and optimizing public transport using cellphone data. *Lecture Note in Computer Science* 8190: 663–666

Calabrese F, Pereira F, Lorenzo G, Liu L, and Ratti C 2010a The geography of taste: Analyzing cell-phone mobility and social events. *Lecture Note in Computer Science* 6030: 22–37

Calabrese F, Reades J, and Ratti C 2010b Eigenplaces: Segmenting space through digital signature. *IEEE Pervasive Computing* 9: 78–84

Calabrese F, Di Lorenzo G, Liu L, and Ratti C 2011 Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Computing* 10: 36–44

Chen J, Shaw S L, Yu H, Lu F, Chai Y, and Jia Q 2011 Exploratory data analysis of activity diary data: A space-time GIS approach. *Journal of Transport Geography* 19: 394–404

Crane, R and Crepeau R 1998 Does neighborhood design influence travel?: A behavioral analysis of travel diary and GIS data. *Transportation Research Part D: Transport and Environment* 3: 225–238

Doherty S T, Noel N, Gosselin M L, Sirois C, and Ueno M 2001 Moving beyond observed outcomes - Integrating global positioning systems and interactive computer-based travel behavior surveys. *Transportation Research Circular*: 449–466

Ferrari L, Mamei M, and Colonna M 2014 Discovering events in the city via mobile network analysis. *Jounral of Ambient Intelligence and Humanized Computing* 5: 265–277

Gao S, Liu Y, Wang Y, and Ma X 2013a Discovering spatial interaction communities from mobile phone data. *Transactions in GIS* 17: 463–481

Gao S, Wang Y, Gao Y, and Liu Y 2013b Understanding urban traffic-flow characteristics: A rethinking of betweenness centrality. *Environment and Planning B: Planning and Design* 40: 135–153

Gazette of the People's Government of Shenzhen Municipality. Issue No. 17, Serial No. 781, March 23, 2012. (available at http://english.sz.gov.cn/gg/201203/t20120329_1836874.htm)

González M C, Hidalgo C A, and Barabási A L 2008 Understanding individual human mobility patterns. *Nature* 453: 779–782

Guo D 2009 Flow mapping and multivariate visualization of large spatial interaction data. *IEEE Transactions on Visualization and Computer Graphics* 15: 1041–1048

Hägerstrand T1970 What about people in regional science? *Papers of the Regional Science Association* 24: 6–21

Jiang S, Fiore G, Yang Y, Ferreira J, Frazzoli E, and González M C 2013 A review of urban computing for mobile phone traces: Current methods, challenges, and opportunities. In *Proceedings of the 2nd SIGKDD Workshop on Urban Computing*. ACM, Chicago, Illinois: 1–9

Kang C, Gao S, Lin X, Xiao Y, Yuan Y, Liu Y, and Ma X 2010 Analyzing and geo-visualizing individual human mobility patterns using mobile call records. In *Proceedings of the 18th International Conference on Geoinformatics.* Beijing, China: 1–7

Kang C, Ma X, Tong D, and Liu Y 2012a Intra-urban human mobility patterns: An urban morphology perspective. *Physica A: Statistical Mechanics and its Applications* 391: 1702–1717

Kang C, Liu Y, Ma X, and Wu L 2012b Towards estimating urban population distributions from mobile call data. *Journal of Urban Technology* 19: 3–21

Kang C, Zhang Y, Ma X, and Liu Y 2013 Inferring properties and revealing geographical impacts of intercity mobile communication network of China using a subnet data set. *International Journal of Geographical Information Science* 27: 431–448

Kloeckl K, Senn O, Di Lorenzo G, and Ratti C 2011 Live Singapore! - An urban platform for real-time data to program the city. In *Proceedings of Computers in Urban Planning and Urban Management (CUPUM 2011)*. Alberta, Canada.

Laney D 2001 3D data management: Controlling data volume, velocity, and variety. Application *Delivery Strategies*. META Group. (available at http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf)

Liao T W 2005 Clustering of time series data - a survey. *Pattern Recogn* 38: 1857–1874

Lin J, Keogh E, Lonardi S, and Chiu B 2003 A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*. ACM, San Diego, California, USA: 2–11

Liu L, Hou A, Birdman A, Ratti C, and Jun Chen 2009 Understanding individual and collective mobility patterns from smart card records: A case study in Shenzhen. In *Proceedings of the 12th International IEEE Conference on Intelligent Transportation Systems*: 1–6

Neuhaus F 2010 *Tweet times - activity over 24 hours in Munich*. (available at http://urbantick.blogspot.com/2010/07/tweet-times-activity-over-24-hours-in.html)

Pei T, Sobolevsky S, Ratti C, Shaw S L, Li T, and Zhou C 2014 A new insight into land use classification based on aggregated mobile phone location data. *International Journal of Geographical Information Science* 28: 1–20

Rae A 2009 From spatial interaction data to spatial interaction information? Geovisualisation and spatial structures of migration from the 2001 UK census. C*omputers, Environment and Urban Systems* 33: 161–178

Ratti C, Pulselli R M, Williams S, and Frenchman D 2006 Mobile landscapes: Using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design* 33: 727–748

Reades J, Calabrese, Sevtsuk A, and Ratti C 2007 Cellular census: Explorations in urban data collection. *IEEE Pervasive Computing* 6: 30–38

Reades J, Calabrese F, and Ratti C 2009 Eigenplaces: analyzing cities using the space-time structure of the mobile phone network. *Environment and Planning B: Planning and Design* 36: 824–836

Reades J 2013 *The MapThing Processing library*. (available at http://www.reades.com/2013/04/01/the-mapthing-processing-library/)

Reas C and Fry B 2010 *Getting Started with Processing*. O'Reilly Media

Schlich R and Axhausen K 2003 Habitual travel behaviour: Evidence from a six-week travel diary. *Transportation* 30: 13–36

Schmid C 2012 *Visualizing urban flows with mobile data*. (available at http://lanyrd.com/2012/uxcon/szmfc/)

Shaw S L, Yu H, and Bombom L S 2008 A space-time GIS approach to exploring large individual-based spatiotemporal datasets. *Transactions in GIS* 12: 425–441

Slingsby A, Beecham R, and Wood J 2013 Visual analysis of social networks in space and time using smartphone logs. *Pervasive and Mobile Computing* 9: 848–864

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

Song C, Koren T, Wang P, and Barabási A L 2010b Modeling the scaling properties of human mobility. *Nature Physics* 6: 818–823

Tan P N, Steinbach M, and Kumar V 2005 *Introduction to Data Mining*. Addison-Wesley

Tobler W R 2007 *Flow Mapper Tutorial*. (available at http://www.csiss.org/clearinghouse/FlowMapper/FlowTutorial.pdf)

Traag V A, Browet A, Calabrese F, and Morlot F 2011 Social event detection in massive mobile phone location data using probabilistic location inference.

In Proceedings of the 2011 IEEE International Conference on Privacy, Risk, and Trust, and IEEE International Conference on Social Computing. Boston, MA, USA: 9–11

Vaccari A, Martino M, Rojas F, and Ratti C 2010 Pulse of the city: Visualizing urban dynamics of special events. In *Proceedings of the 20th International Conference on Computer Graphics and Vision*. ACM, New York, NY: 64–71

Viégas F and Wattenberg M 2012 Wind map. (available at http://hint.fm/wind/)

Wolf J, Guensler R, and Bachman W 2001 Elimination of the travel diary: Experiment to derive trip purpose from global positioning system travel data. In *Proceedings of the Transportation Research Board 80th Annual Meeting*. Washington, DC, USA: 125–134

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

Yuan J, Zheng Y, and Xie X 2012a Discovering regions of different functions in a city using human mobility and POIs. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Beijing, China: 186–194

Yuan Y and Raubal M 2012 Extracting dynamic urban mobility patterns from mobile phone data. *Lecture Note in Computer Science* 7478: 354–367

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

Zheng Y, Capra L, Wolfson O, and Yang H 2014 Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5: 1–55

Zhu X and Guo D 2014 Mapping large spatial flow data with hierarchical clustering. *Transactions in GIS* 18: 421–435

Chapter 3

Understanding the Bias of Call Detail Records in Human Mobility Research

Abstract

In recent years, call detail records (CDRs) have been widely studied in human mobility research. Although CDRs are originally collected for billing purposes, the vast amount of digital footprints generated by calling and texting activities provide useful insights about population movement. However, can we fully trust CDRs given the uneven distribution of people's phone communication activities in space and time? In this paper, we investigate this issue using a mobile phone location dataset collected from over one million subscribers in Shanghai, China. It includes CDRs (~27%) plus other cellphone-related logs (e.g., tower pings, cellular handovers) generated in a workday. We extract all CDRs in a separate dataset in order to compare human mobility patterns derived from CDRs and from the complete dataset. From an individual perspective, the effectiveness of CDRs in estimating three frequently used mobility indicators is evaluated. We find that CDRs tend to underestimate total travel distance and movement entropy, while they can provide a close estimate to radius of gyration. In addition, we learn that the level of deviation is also relevant to the ratio of CDRs in one's trajectory. From a collective perspective, we compare outcomes of two datasets in distance decay effect analysis and urban community detection. The major differences can be explained by the habit of mobile phone usage in space and time. We believe the event-triggered nature of CDRs does introduce certain degree of bias in human mobility research and we suggest people use caution in future research.

3.1 Introduction

The advent of the so-called "big data era" offers many new opportunities to resolve the mystery of human mobility using various types of massive digital footprints, such as geo-tagged social media data (Batty 2010). Despite those exciting discoveries that reveal the pulse of the city, there have been debates regarding the biases that come with the data. For instance, studies report that distribution of social media users is predominantly uneven in terms of geography, gender, and race/ethnicity (Mislove *et al.* 2011, Hecht and Stephens 2014).

Mobile phone location data, collected by mobile network operators (MNOs), also has been an appealing data source, given the unprecedented scale of digital footprints it carries. The type of mobile phone location data used by most existing studies is referred to as call detail records (CDRs), which are generated upon phone communication activities (i.e., make/receive a phone call, send/receive a text message). For billing purposes, CDRs keep track of related information (e.g., caller/callee, time, duration) of each event, plus the unique identifier of the nearby cell tower that handles the phone communication.

Numerous valuable findings regarding human activity and urban environment, as well as their interactions, have been uncovered since CDRs became prevalent in the research community in recent years (e.g., González *et al.* 2008, Song *et al.*

2010a, Song *et al.* 2010b). However, the majority of previous studies do not mention how representative their data (i.e., CDRs) are, as well as the applicability of analysis outcomes to the entire population. Also, very few researchers study what CDRs cannot do (see Kang *et al.* 2012 for an example). Are we overly optimistic about the usefulness of CDRs and the validness of our conclusions? Like what has been brought up, debated, and acknowledged in the social media community, the representativeness of CDRs needs to be carefully examined.

As pointed out by Becker *et al.* (2013), CDRs are coarse in space and sparse in time. In large cities, the spatial granularity at the cell tower level may not be a major drawback as cell towers are usually densely distributed all over the urban area. What really matters is the uneven distribution of people's phone communication activities in space and time. On one hand, people are more likely to contact others at certain places, such as home or work location, and it is highly possible that those locations account for only a fraction of all visited places. On the other hand, depending on how actively one engages in phone communication, the total number of CDRs each subscriber generates varies significantly. The dataset used in this research reveals that population size drops with the increased intensity of phone-related activities (Figure 3.1). Around 17% subscribers in our dataset have two or less CDRs in a day and over 38% subscribers generate less than seven CDRs. Hence, whether mobility pattern of subscribers without heavy phone usage can be characterized is indeed questionable. One may argue that this problem can be solved by collecting CDRs over a period of time, such as a week, a month, or even longer. Although this workaround does help increase sample size, the uneven spatiotemporal distribution of digital footprints caused by people's habit of mobile phone usage cannot be addressed. The "quiet minority" who rarely make use of mobile device remain underrepresented.

This research takes the first step to evaluate the representativeness of CDRs in human mobility characterization, using a mobile phone location dataset that includes both CDRs and non-CDR footprints. The latter are generated by events irrelevant to phone communication, such as moving out of the service area of a cell tower, active pinging from cell tower, and so forth. By isolating the CDR part in a separate dataset, we are able to quantitatively evaluate the effectiveness of CDRs in human mobility analysis, from both the individual perspective and the collective perspective. The findings of this research not only facilitate a better understanding of CDRs as a remarkable data source, but also lead us to rethink of some existing findings of human mobility researchers have come up with so far.

The remainder of this paper is organized as follows. The next section discusses existing research related to this study. Section 3.3 presents the study area and the mobile phone location data used in this research. In Section 3.4, we adopt an individual perspective and evaluate the effectiveness of CDRs in estimating some of the most frequently used mobility indicators. We then take a collective approach

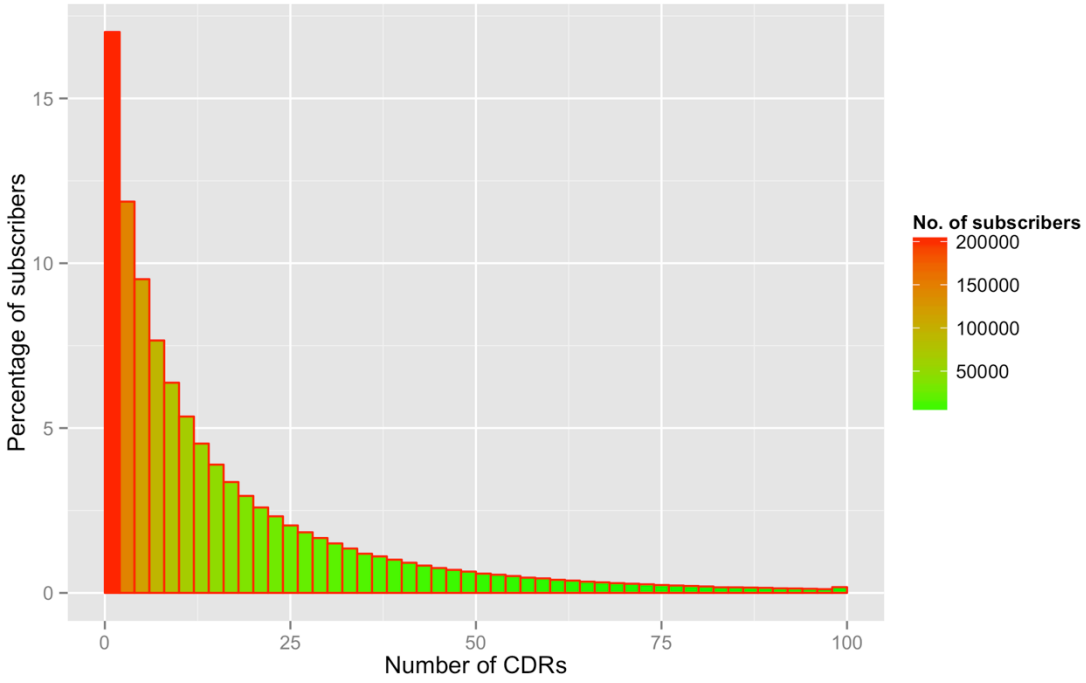

Figure 3.1 Distribution of subscribers under different intensity levels of phone communication.

in Section 3.5 and examine the performance of CDRs in distance decay effect analysis and urban community detection. We conclude and discuss this research in Section 3.6.

3.2 Relevant research

This section discusses relevant research in the following two areas: 1) CDRs and human mobility; 2) CDRs and urban dynamics; and 3) uncertainty issue.

3.2.1 CDRs and human mobility

Thanks to CDRs, our knowledge of individual human mobility has been enriched considerably in recent years. A large body of literature focuses on individual activity space, which is used to denote the spatial configuration for people's daily activities (Golledge and Stimson 1997). Understanding individual space has a profound meaning in the real world, such as accessibility to healthcare facilities (Sherman *et al*. 2005), environmental exposure (Perchoux *et al*. 2013), etc. Note that the term activity space is related to several other concepts, for instance, awareness space (Brown and Moore 1970), action space (Horton and Reynold 1971), space-time prism (Hägerstrand 1970).

Activity space can be characterized from individual trajectories, as each of them reflects a person's movement in space over time. A number of measures can be used to describe the spatiotemporal characteristics of individual trajectory, such as daily range of travel, movement radius, movement entropy (e.g., Yuan *et al.* 2012). Two dimensional measures, such as standard deviational ellipse (SDE), can be used to describe the range and direction of one's activity space (e.g., Zenk *et al*. 2011). Based on these mobility indicators, statistical analyses can be performed to compare activity space of people in different social groups (e.g., age, gender, see Kang *et al.* 2010, Yuan *et al.* 2012) or people at different locations (Becker *et al.* 2013). In addition, individual trajectories collected over a long term allow us to extract meaningful anchor points of one's activity space, such as home or work locations (e.g., Ahas *et al.* 2010, Calabrese *et al.* 2010a), and to examine people's activity patterns around anchor points (e.g., Xu *et al.* 2015).

Besides activity space research, CDRs have been utilized by physicists to gain new insights on the nature of human travel. Over a long period of time, we believe that human movements are associated with a large degree of randomness and can thus be explained by the random walk or Lévy flight model (Brockmann *et al.* 2006, Rhee *et al.* 2011). However, analysis results from CDR trajectories have proved that human travels actually follow reproducible patterns (González *et al.* 2008) and are highly predictable (Song *et al.* 2010a, Song *et al.* 2010b).

Despite substantial progress discussed in this section, this study argues that the representativeness of individual trajectory derived from CDRs is strongly subjected to people's habit of mobile phone usage in space and time. The CDR trajectory of a traveling salesperson who keeps talking to customers by phone can well depict his/her daily travel, whereas the CDR trajectory of someone who only contacts parents once a week should not be used to understand his/her mobility pattern in space and time. As a result, can we fully trust those mobility indicators derived from CDR trajectories, as well as other relevant conclusions? It should be pointed out that using CDRs collected over a long period of time as a workaround cannot address this problem as people who rarely engage in phone communication remain underrepresented.

3.2.2 CDRs and urban dynamics

Instead of focusing on individual trajectories, many studies adopt a collective approach to uncover varying mobility patterns by location. Frequently used indicators include Erlang value (i.e., the total call traffic volume in one hour), number of phone calls/text messages, number of active subscribers, etc. For instance, CDRs are used to quantitatively measure different levels of popularity in New York City in terms of the density and distribution of aggregate phone calls (Girardin *et al.* 2009). Distinct patterns of mobility variation throughout different time periods in a day, or different days in a week can also be extracted and compared using a variety of techniques, such as K-Means (Reades *et al.* 2007),

eigendecomposition (Calabrese *et al.* 2010b), dynamic time warping (Yuan and Raubal 2012). In addition to mobility pattern analysis, aggregate population flows among cell towers serve as an indication of human interaction with urban space, which enable us to detect urban communities with strong internal interactions (Gao *et al.* 2013). Moreover, some recent studies utilize the characteristics of people's phone communication activities and develop innovative methodologies to address problems that are usually solved by other approaches. For example, Pei *et al.* (2014) develop a new method for urban land use classification based on normalized hourly call volume and the total call volume.

Similar to human mobility research, many CDR-based urban dynamics studies also make a fundamental assumption that phone communication records can serve as a direct indication of human activity intensity, which in fact is debatable. For instance, can phone calls be considered as "a proxy for presence of people" (Girardin *et al.* 2009)? A careful evaluation of the representativeness of CDRs can help us answer this type of question.

3.2.3 Uncertainty issue

Uncertainty has been an important topic in GIScience (Goodchild and Gopal 1989, Zhang and Goodchild 2002). It is associated with a series of concepts, such as accuracy, precision, consistency, completeness, to name a few (Veregin 1999). Considerable efforts have been made to visualize and analyze spatiotemporal uncertainties (Pang 2001, MacEachren *et al*. 2005, Delmelle *et al.* 2014). With an improved understanding of uncertainties, many critical concerns have been raised regarding how the uncertainties could influence our knowledge (e.g., Griffith *et al.* 2007, Zinszer *et al*. 2010, Jacquez 2012), as well as the risk in the decision making process (Golledge and Stimon 1997).

The issue of uncertainty is mainly examined in the field of environmental modeling (Refsgaard *et al*. 2007, Ascough II *et al*. 2008). Despite limited discussion in literature, uncertainties that come with mobile phone location data due to the relatively coarse spatiotemporal granularity should not be ignored. From the spatial perspective, the resolution of spatial location is restricted at the cell tower level (Becker *et al*. 2013). In urban area where cell towers are sparsely distributed, the distance from a cell tower to the closest one can be longer than 1 km. From the temporal perspective, the location of a subscriber between two phone communication events is uncertain. Provided with a two-hour interval, the potential area that a subscriber can travel to may cover the entire city. With CDR data, the duration between phone communication activities is often longer than 2 hours, which leads to a large degree of uncertainty in human mobility analysis. Note that the uncertainties that result from coarse spatial granularity cannot be overcome because of the fixed number and distribution of cell towers, whereas the temporal granularity of individual footprints can be improved. Instead of enforcing

subscribers to make more phone calls, different collection methods, such as active pinging, can help reduce the uncertainties between two consecutive footprints. The uncertainty issue itself is not a nightmare and the more important matter is to understand the how uncertainties can result in imperfect knowledge and recognize "which cannot be known" (Couclelis 2003). This is the fundamental objective of this paper.

3.3 Data

Our study area is Shanghai, one of the largest cities in China. In this section, we introduce some background information of Shanghai and the mobile phone location dataset collected in this city.

3.3.1 Area of study

Shanghai has a resident population of 23.8 million as of 2012 (Shanghai Municipal Statistics Bureau, 2012), which makes it the largest city in China by population. Shanghai is one of the global financial centers and the busiest container port in the world (World Shipping Council, 2013). Its annual gross domestic product (GDP) also ranks No.1 in China in 2012 (National Bureau of Statistics of China, 2012).

Located in the central east coast of China, Shanghai has a total area of 6,340.5 square kilometers (Shanghai Municipal Statistics Bureau, 2012). It consists of 16 administrative districts and the Chongming County (Figure 3.2). Among those districts, eight of them on the west bank of the Huangpu River (Huangpu, Xuhui, Jingan, Changning, Yangpu, Hongkou, Putuo, and Zhabei), also known as Puxi, are referred to as the downtown area of Shanghai. Over the past two decades, the economy of the Pudong District, situated on the east bank of the Huangpu River, has been growing rapidly, with its famous zone of Lujiazui being widely considered as the financial center of Shanghai.

3.3.2 Dataset

The mobile phone location dataset used in this study is collected by a major MNO in China. It is obtained through a joint research collaboration. It includes all records generated by 1,252,797 subscribers on September 3, 2012. Different from those analyzed by previous studies, this dataset contains both CDRs and actively generated logs, differentiated by seven event codes listed in Table 3.1. This particular MNO owns 33,044 cell towers all over Shanghai and the cell tower ID associated with each record indicates approximate location where each event takes place. It should be pointed out that to protect individual privacy, we do not possess any personal information (e.g., age, gender, phone number) and the spatial granularity is restricted at cell tower level.

Figure 3.3 demonstrates the total number of each event recorded during every hour. Given the way events are triggered, those numbers vary differently

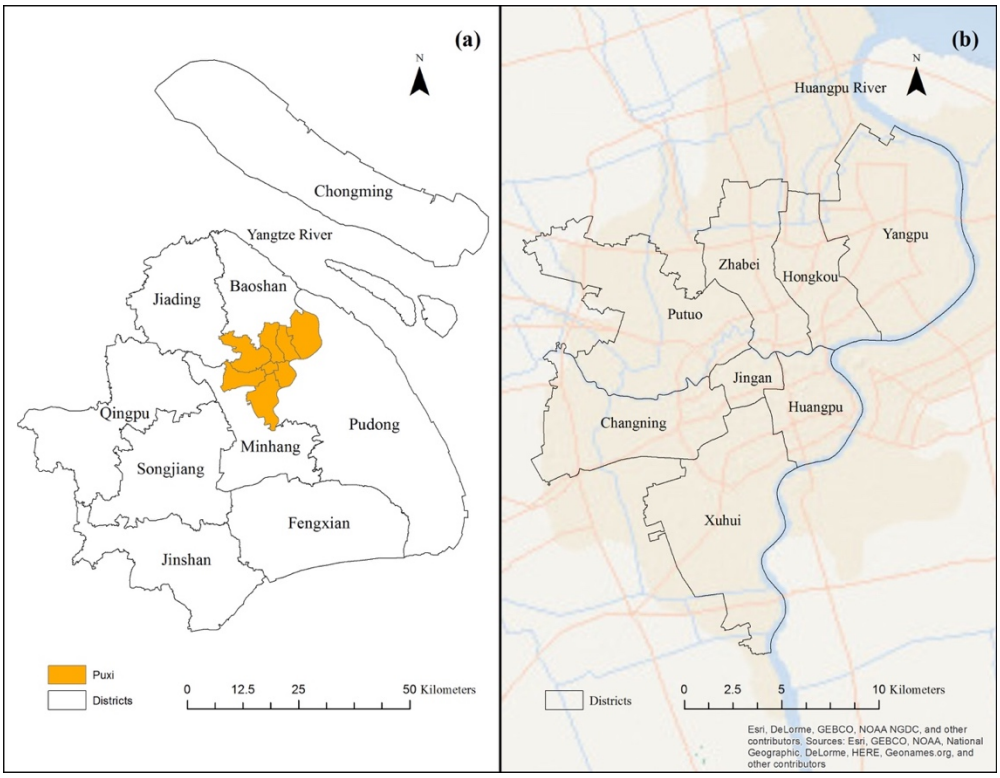

Figure 3.2 (a) Shanghai and its administrative districts. The orange areas represent the "Puxi" region, the downtown area of Shanghai. (b) Eight administrative districts in the "Puxi" region. "Puxi" and the Pudong districts are divided by the Huangpu River.

throughout the day. In general, except PU, very few records are generated during midnight. The city "wakes up" at 6:00 – 7:00, indicated by significantly elevated numbers of generated records. Again, PU is an exception because of the growth of RU and CH triggered by enhanced human mobility. As a result, the number of active pinging from tower, recorded as PU, declines accordingly. The peaks of RU at 8:00 – 9:00 and 17:00 – 18:00 correspond to the morning and evening rush hour, respectively. Similar to RU, the numbers of IN and OT events start to increase from 6:00 – 7:00. ON and OF events together account for a very small portion of the data as turning mobile phone on and off repeatedly is not a common practice.

3.3.3 Data processing

The various types of events recorded in this dataset offer a valuable opportunity of understand the bias of CDRs in human mobility analysis. For the purposes of direct comparison, we extract all CDRs (i.e., IN and OT events) from every subscriber and store them in a separate dataset. Therefore, each subscriber has two groups of data: CDRs and the entire set of records. In the remainder of this paper, we call these two datasets as the CDR group and the complete group, respectively.

Table 3.1 Summary of event codes.

Code	Event	Description	Avg. Record No. Per Subscriber
RU	Regular update	Regular update triggered by moving from the service area of a cell tower to that of another tower.	12.51
PU	Periodic update	Periodic update triggered by tower pinging if subscriber has been "silent" (no other events in this table are detected) for a certain time period. However, the specific criteria (e.g., duration of silence) for triggering a periodic update is not clear. Moreover, mobile phones which are turned off or disconnected from the cellular network cannot receive tower pining.	4.88
OT	Phone communication (outbound)	Subscriber makes a phone call or sends a text message.	4.45
ON	Power on	Mobile phone is turned on and connected to cellular network.	0.62
OF	Power off	Mobile phone is turned off and disconnected from cellular network.	0.39
IN	Phone communication (inbound)	Subscriber receives a phone call or a text message.	14.67
CH	Cellular handover	Transfer of an ongoing phone call from one cell tower to another caused by movement.	5.45

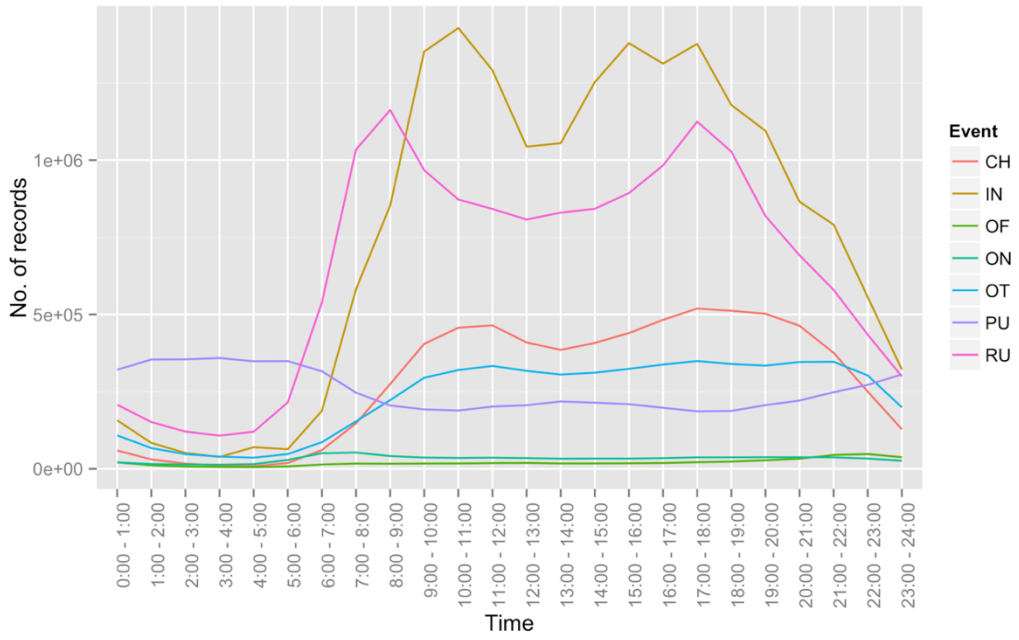

Figure 3.3 Temporal variation of the total number of each event.

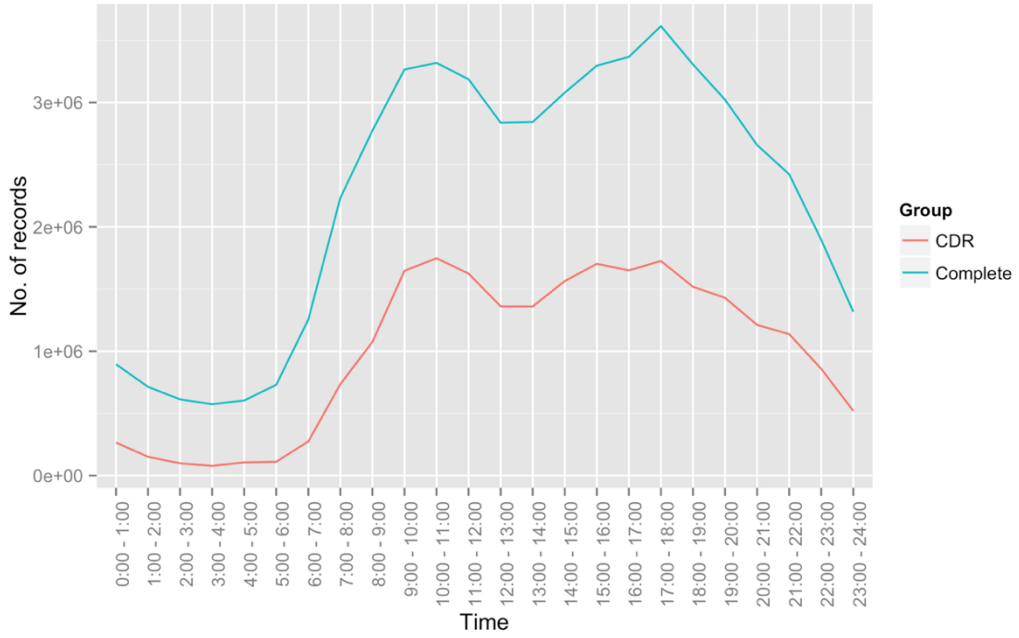

Figure 3.4 Temporal variation of the total number of records in the CDR group and the complete group.

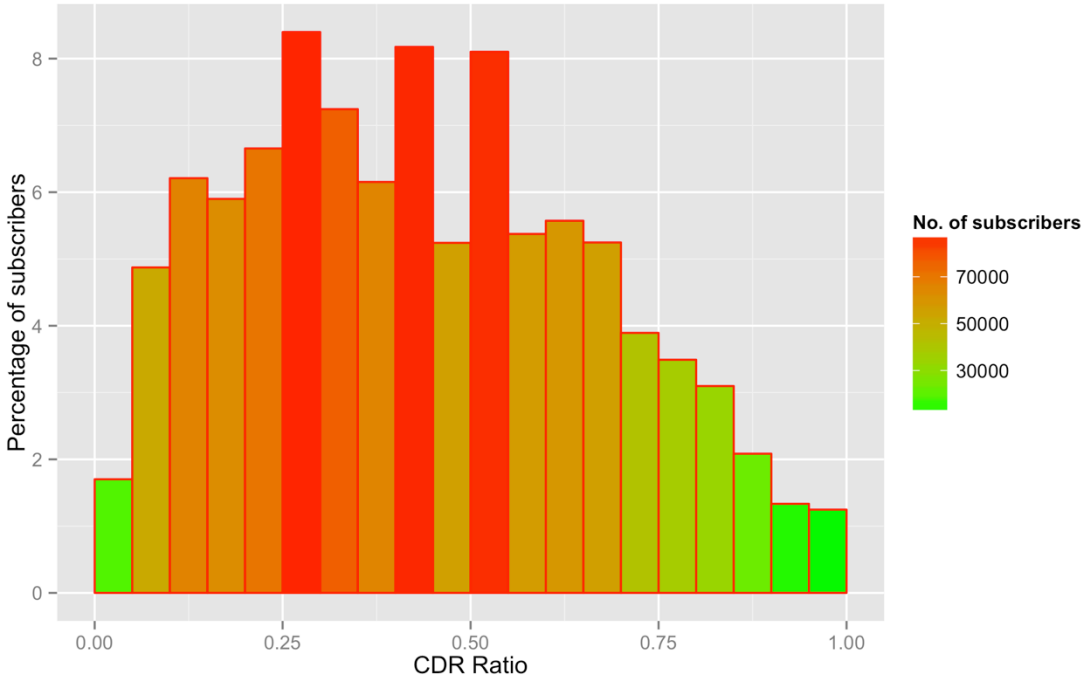

Figure 3.5 Distribution of subscribers under different ratio of CDRs.

As a subset of the data, the temporal variation of the total number of records in the CDR group mirrors that in the complete group (Figure 3.4), although the former does not reveal a striking upsurge during 17:00 – 18:00. For each subscriber, the CDR ratio (i.e., number of CDRs / number of total record) has a mean of 43.09% and a median of 41.18%. However, depending on how actively one engages in phone communication activities, this number varies significantly (Figure 3.5).

3.4 Individual human mobility

This section focuses on evaluating the representativeness of CDRs in individual daily mobility pattern analysis. We aim to answer the following question: compared with the complete set of footprints, how well do CDR footprints characterize one's daily mobility pattern? We focus on three basic properties of human mobility: distance, range, and heterogeneity. Hence, the following three frequently used mobility indicators are evaluated quantitatively: 1) total travel distance; 2) radius of gyration; and 3) movement entropy. In the evaluation process, the complete group is regarded as the control group and mobility indicators derived from the complete group are considered to be more accurate. To better address our research question, two additional data processing steps are performed.

First, we need be certain that records in the complete group can characterize one's daily mobility pattern to a good extent. If all footprints generated by a subscriber

in a day cannot provide sufficient temporal coverage (e.g., people who keep their mobile phone turned off most of the time), this subscriber's daily mobility pattern would remain a mystery so the complete set of footprints cannot be used as a benchmark to evaluate the representativeness of his/her CDR footprints. To reduce the level of uncertainty, we divide the day into four six-hour periods (0:00 – 6:00, 6:00 – 12:00, 12:00 – 18:00, and 18:00 – 24:00) and only those subscribers with at least one footprint in each six-hour period are eligible for the proposed evaluation. After this step, a total of 686,642 subscribers are selected.

Second, as discussed in Section 3.3.3, the range of CDR ratio varies significantly among subscribers and the CDR ratio could be a critical factor in the evaluation process. If one's footprints are mostly generated by phone communication activities, mobility indicators derived from the CDR group and the complete group should be very close. On the contrary, for people who travel a lot but rarely contact others, his/her CDRs are likely to yield very biased mobility indicators. In order to understand how the CDR ratio influences the estimation of mobility indicators, we further break down those 686,642 subscribers into four classes by CDR ratio (Table 3.2).

Table 3.2 Summary of four subscriber classes divided by CDR ratio.

Class	CDR Ratio	Number of Subscribers
A	75% – 100%	60,519
B	50% – 75%	173,940
C	25% – 50%	251,187
D	0% – 25%	200,996

3.4.1 Total travel distance

Total travel distance is the aggregated length of one's daily movement and it is a basic measure of individual mobility. It is calculated as the sum of Euclidian distance between each pair of consecutive footprints. For each subscriber, we compute two values of total travel distance, D_{cdr} and $D_{complete}$, based on the CDR group and the complete group. Results from all subscribers are plotted on a two-dimensional space (Figure 3.6). The horizontal axis and the vertical axis represent the complete group and the CDR group, respectively. In this figure, the horizontal axis is binned with a bandwidth of 0.1 km. Subscribers are aggregated in terms of 1) class assignment on the basis of CDR ratio (Table 3.2), and 2) which 0.1-km bin $D_{complete}$ falls in. Then, for aggregated subscribers in each bin, the average value of D_{cdr} is computed and plotted. This diagram allows us to examine the representativeness of CDRs by visual inspection: if CDRs are representative, points on Figure 3.6 should be close to the diagonal. On the contrary, a large deviation from the diagonal leads to the conclusion that CDRs tend to

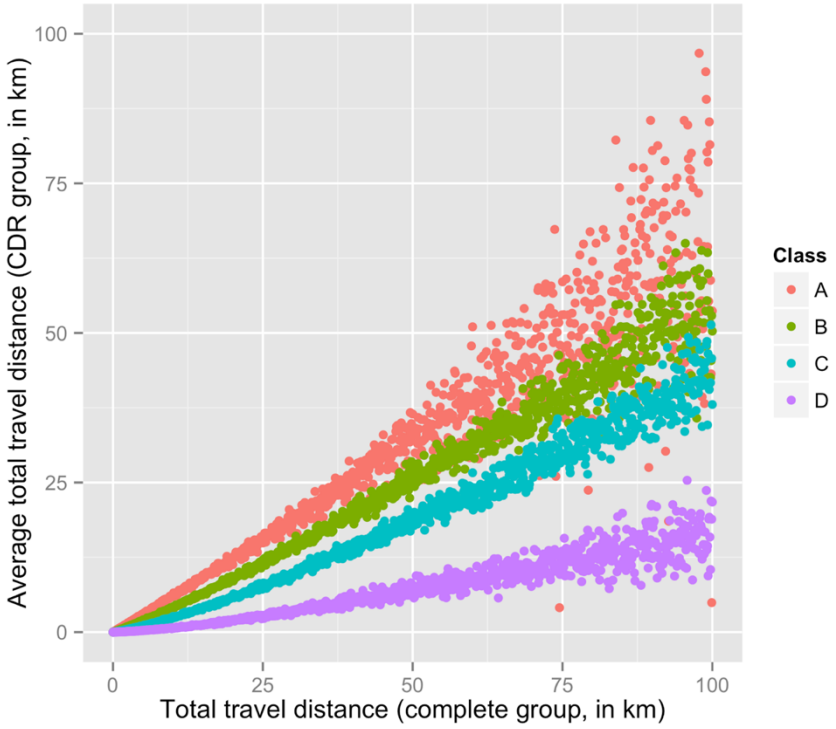

Figure 3.6 Total travel distance (complete group) vs. average total travel distance (CDR group).

underestimate total travel distance. For this mobility indicator, overestimation is not possible given that D_{cdr} cannot be greater than $D_{complete}$.

Several interesting findings regarding the effectiveness of CDRs in estimating total travel distance are revealed in Figure 3.6. First, all four classes suggest that D_{cdr} and $D_{complete}$ have very high positive correlation. This is confirmed by both Pearson correlation coefficient and Spearman correlation coefficient (Table 3.3). Second, CDR ratio does matter. As the CDR ratio declines (Class A -> Class D), points in Figure 3.6 deviates more from the diagonal. To quantify the level of underestimation, we fit points in each subscriber class with a linear regression model: $y = \alpha + \beta x$, using $D_{complete}$ as the independent variable x and D_{cdr} as the dependent variable y. The regression coefficient β indicates the relationship between D_{cdr} and $D_{complete}$, while $1 - \beta$ can be interpreted as the level of underestimation, which implies how well CDRs can estimate one's total travel distance. It is evident that CDRs tend to significantly underestimates total travel distance even for subscribers in Class A, whose CDRs account for at least 75% of all footprints (Table 3.4). On average, D_{cdr} of a Class A subscriber is 34.3% shorter than his/her $D_{complete}$. This regression coefficient turns out to be smaller in Class B and Class C, which indicates that CDRs become more and more biased

Table 3.3 Correlation between total travel distance (complete group) and average total travel distance (CDR group).

Class	Pearson correlation	Spearman correlation
A	0.941	0.960
B	0.987	0.992
C	0.989	0.994
D	0.955	0.976

Table 3.4 Linear regression results between total travel distance (complete group) and average total travel distance (CDR group).

Class	Regression coefficient (β)	Level of underestimation $(1-\beta)\%$
A	0.657	34.3%
B	0.556	44.4%
C	0.443	55.7%
D	0.174	82.6%

in estimating total travel distance if CDR ratio drops. For subscribers in Class D, their CDRs on average underestimate total travel distance by 82.6%. Figure 3.6 also suggests that the variation of average D_{cdr} becomes larger when the value of $D_{complete}$ grows. It is not difficult to make sense out of this pattern of heteroscedasticity under the context of human travel: if one's daily travel distance is longer, the range of estimated travel distance based on his/her CDRs is expected to be wider. Another possible reason is that the size of subscribers drops rapidly when the total travel distance increases. It may also results in a wider range of average D_{cdr}.

3.4.2 Radius of gyration

Radius of gyration is one of the most frequently used measures of activity space. It is defined as the root mean squared distance between a set of visited locations up to time t and the center of mass:

$$r_g^a(t) = \sqrt{\frac{1}{n_c^a(t)}\sum_{i=1}^{n_c^a}(\overrightarrow{r_i^a} - \overrightarrow{r_{cm}^a})^2} \quad (1)$$

where $\overrightarrow{r_i^a}$ represents the $i = 1, \dots, n_c^a(t)$ location of subscriber a and $\overrightarrow{r_{cm}^a} = \frac{1}{n_c^a(t)}\sum_{i=1}^{n_c^a}\overrightarrow{r_i^c}$ defines the center of mass (González *et al.* 2008). Radius of gyration reflects the range of activity space, typically around the center of home and work locations for commuters.

Similar to Section 3.4.1, we compute two values of radius of gyration for each subscriber based on the CDR group and the complete group, denoted as R_{cdr} and

$R_{complete}$. Figure 3.7 uses the horizontal axis to represent the complete group with a 0.1-km bandwidth and the vertical axis to represent the average R_{cdr} of subscribers in the same 0.1-km bin. Again, the consistency between two groups can be inferred by the closeness of data points to the diagonal. Note that unlike total travel distance, R_{cdr} might be larger than $R_{complete}$, if CDR footprints are spread more widely than non-CDR ones.

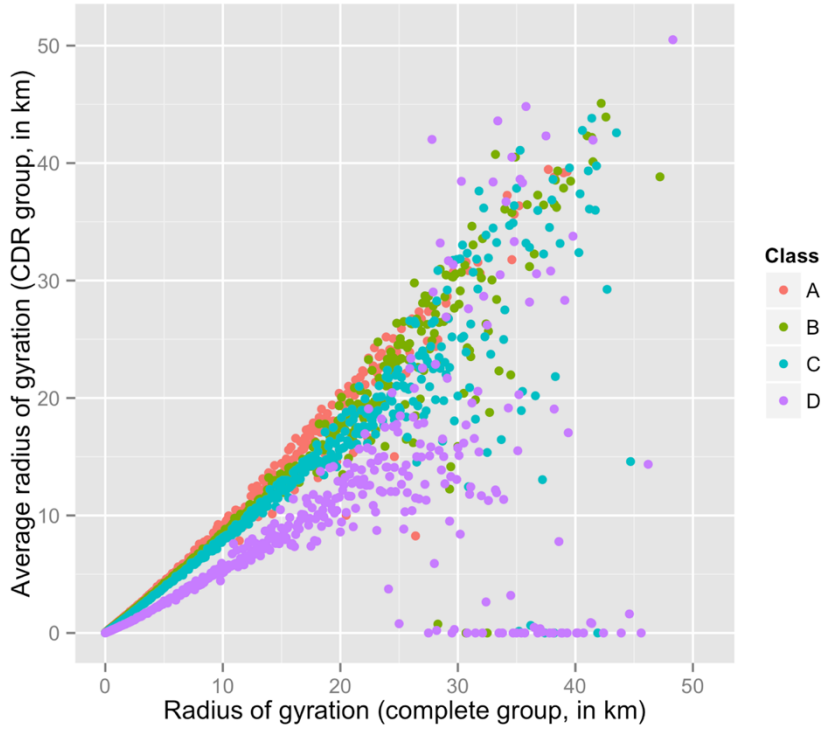

Figure 3.7 Radius of gyration (complete group) vs. average radius of gyration (CDR group).

The effectiveness of CDRs in estimating radius of gyration is noteworthy. First, For Class A, B, and C, R_{cdr} are strongly correlated with $R_{complete}$ (Table 3.5). However, both Pearson correlation coefficient and Spearman correlation coefficient show a significant drop of correlation in Class D, although they still suggest a positive correlation. In addition, for Class D subscribers whose $R_{complete}$ is larger than 25 km, their average values of R_{cdr} are often zero, or very close to zero. Therefore, CDRs might significantly underestimate the radius of gyration of people who commute over long distance and rarely use mobile phone. Second, Figure 3.7 reveals a pattern of heteroscedasticity: the range of estimated radius of gyration based on CDRs is supposed to be wider if one's radius of gyration derived from the complete group grows. From Class A to Class D, such pattern of

Table 3.5 Correlation between radius of gyration (complete group) and average radius of gyration (CDR group).

Class	Pearson correlation	Spearman correlation
A	0.980	0.982
B	0.936	0.939
C	0.860	0.882
D	0.532	0.521

heteroscedasticity turns out to be more obvious as CDRs become a smaller part of one's footprints.

By fitting a linear model to each class we can quantify the effectiveness of CDRs for estimating this mobility indicator. The regression coefficient is very high for Class A and B (> 90%, see Table 3.6). It means that CDRs could depict the range of daily travel very well for subscribers whose 50% or more footprints are collected by phone communication events. Taking into account other non-CDR footprints makes very limited difference on the derived radius of gyration. For subscribers in Class C, CDRs on average underestimate their radius of gyration by 22.4%. Depending on specific applications, this margin of error may be acceptable. However, the small regression coefficient (0.423) in Class D indicates that CDRs fail to provide a good estimate for subscribers whose fraction of CDRs is below 25%. Many subscribers in this group engage none, or very few phone communications in a day. Others who make good use of mobile phones also travel a lot and leave numerous digital footprints (RU event) in the meantime. As a result, those CDRs remain insufficient for deriving daily activity space.

Table 3.6 Linear regression results between radius of gyration (complete group) and average radius of gyration (CDR group).

Class	Regression coefficient (β)	Level of underestimation $(1-\beta)$%
A	0.979	2.1%
B	0.915	8.5%
C	0.776	22.4%
D	0.423	57.7%

3.4.3 Movement entropy

Movement entropy measures the heterogeneity of visitation patterns (Song *et al.* 2010, Yuan *et al.* 2012). It can be calculated using the following equation:

$$E = -\sum_{i=1}^{n} p_i log_2 p_i \quad (2)$$

where n is the number of distinct locations (i.e., cell towers) visited by subscriber and p_i is the probability that location i is visited. Mathematically, the value of

movement entropy grows with a more heterogeneous visitation pattern. Consider the following examples:

1) If a subscriber stays at a single location, $E = -(1.0 \times log_2 1.0) = 0$;
2) If a subscriber visits Location A one times and Location B four times, $E = -(0.2 \times log_2 0.2 + 0.8 \times log_2 0.8) \approx 0.72$;
3) If a subscriber visits Location A five times and Location B five times, $E = -(0.5 \times log_2 0.5 + 0.5 \times log_2 0.5) = 1$;
4) If a subscriber visits Location A, B, C, and D, two times each, $E = -(0.25 \times log_2 0.25 + 0.25 \times log_2 0.25 + 0.25 \times log_2 0.25 + 0.25 \times log_2 0.25) = 2$.

For this mobility indicator, E_{cdr} and $E_{complete}$ are calculated for each subscriber. The correlation tests all indicate a very positive correlation between E_{cdr} and $E_{complete}$ (Table 3.7). Unlike the other two selected mobility indicators, we cannot identify an evident pattern of heteroscedasticity when $E_{complete} < 5$ (Figure 3.8). We find that CDRs can estimate the movement entropy very well for subscribers in Class A given the high regression coefficient (0.925, see Table 3.8). This coefficient declines for Class B and Class C, which implies that as the CDR ratio decreases, it is more likely that some non-CDR footprints are collected at other visited locations where subscribers do not engage phone communications. This might be the most reasonable explanation for the low regression coefficient (0.348) associated with Class D. Apparently, CDRs underestimate the movement entropy by far (65.2%) for subscribers in this class. Moreover, data points in Class D suggest some abnormal drop of average E_{cdr} when $E_{complete} > 6$ (Figure 3.8), which is not the case for the other three classes. We belive that it is also caused by the low likelihood of making phone communications at visited locations.

Table 3.7 Correlation between movement entropy (complete group) and average movement entropy (CDR group).

Class	Pearson correlation	Spearman correlation
A	0.998	0.999
B	0.996	0.999
C	0.991	0.997
D	0.900	0.934

Table 3.8 Linear regression results between movement entropy (complete group) and average movement entropy (CDR group).

Class	Regression coefficient (β)	Level of underestimation ($1 - \beta$)%
A	0.925	7.5%
B	0.815	18.5%
C	0.690	31.0%
D	0.348	65.2%

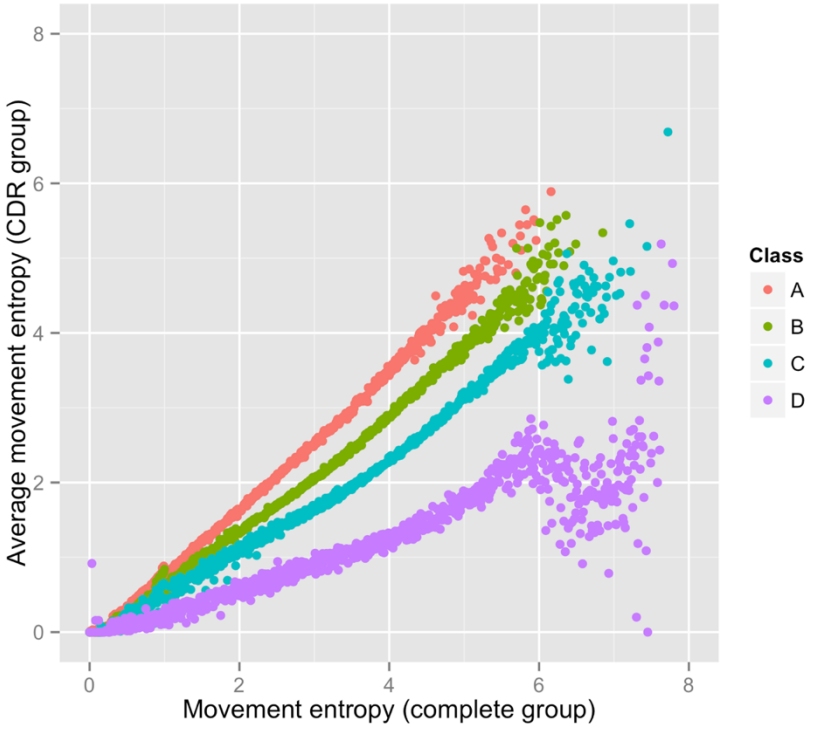

Figure 3.8 Movement entropy (complete group) vs. average movement entropy (CDR group).

In this section, we evaluate the representativeness of CDRs based on mobility indicators that measure activity space from three aspects: distance, range, and heterogeneity. We reveal some fundamental findings by answering "whether CDRs can provide a good estimate of individual mobility patterns". We learn that the answer is not simply yes or no. Perhaps the question should instead be framed as, "how good are CDRs in providing a good estimate of individual mobility patterns". According to our analysis, the effectiveness of CDRs in individual mobility study depends on the research question and the mobility measure selected to address that question. To estimate radius of gyration, CDRs in most cases are probably good enough for subscribers who 1) make at least some phone communications throughout the day, and 2) travel within normal daily activity range (e.g., less than 25 km in Shenzhen). On the contrary, one needs to be cautious when using CDRs to study some problems, such as travel distance, or heterogeneity of human mobility. To a large extent, the validness of analysis result is subject to how actively subscribers engage in phone communications. Therefore, in many cases we might bear the risk of underestimating mobility indicators of interest.

3.5 Collective human mobility

Many researchers approach human mobility study from a collective perspective and pay more attention to data aggregated from individual level. In this section, we evaluate the representativeness of CDRs from this perspective. The analyses of distance decay effect urban community are selected for the evaluation process because they are examined by several CDR-based human mobility studies (e.g., González *et al.* 2008, Walsh and Pozdnoukhov 2011, Gao *et al*. 2013).

3.5.1 Distance decay effect

Existing studies reveal that human motion can be modeled by a Lévy flight (Brockmann *et al.* 2006), while the power law distribution of step lengths is an indication of distance decay effect (Liu *et al*. 2012, Gao *et al*. 2013). The notion of distance decay has a close relationship with *The First Law of Geography*: "everything is related to everything else, but near things are more related than distant things" (Tobler 1970). Nowadays, although in many regions, highly developed urban infrastructure can offer various means of transportation, human activity remains to be restricted by a number of factors, such as distance and accessibility. Many researchers argue that the "death of distance" hypothesis is premature (Wang *et al.* 2003; Rietveld and Vickerman 2004).

The massive collection of CDRs offers some new insights to validate and/or calibrate our understanding of the friction effect of distance. For instance, González *et al.* (2008) and Gao *et al.* (2013) report distance decay parameters of 1.75 and 1.60, respectively. However, as discussed earlier, CDRs are generated only upon phone communication activities and most people do not use their mobile phone at all places they visit. Therefore, displacements between CDR footprints can only represent movements between phone communications. Taking advantages from the various event types recorded in this dataset, we are able to compare the distance decay effect observed from people's phone communication activities with that derived from the complete set of footprints.

We capture 4,992,719 displacements from the CDR group and 27,686,129 displacements from the complete group. Figure 3.9 shows the cumulative distribution function (CDF) plot of two data groups. Both curved lines indicate that most displacements are short: around 90% displacements in the CDR group are below 5 km and roughly 90% of displacements in the complete group are under 2.5 km. Those displacements can be approximated by a power law distribution in the following form:

$$P(d) \propto d^{\beta} \quad (3)$$

where β is the distance decay parameter (Gao *et al.* 2013). A large value of β indicates that distance is a strong deterrent to interaction, whereas a small value of β implies a relatively weak influence of distance.

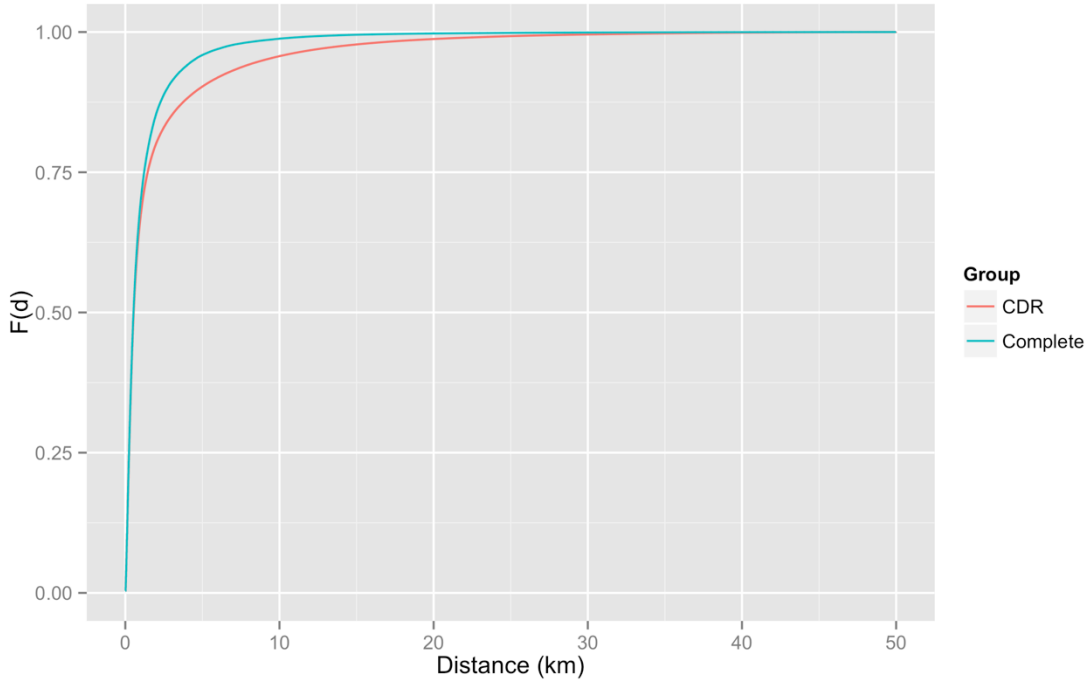

Figure 3.9 Cumulative distribution function (CDF) of displacements.

Figure 3.10 shows the probability density function (PDF) in log-log scale and the fitted power law distribution. The fitted decay parameters are $\beta_1 = 1.79$ for the CDR group and $\beta_2 = 1.98$ for the complete group. Note that β_1 is very close to 1.75, reported by González *et al.* (2008), which indicates that a similar mechanism that drives the friction effect of distance is captured. As what we would expect based on the CDF plot (Figure 3.9), β_2 is larger than β_1 given that the complete group captures more short-distance displacement. For this reason, we believe that CDRs slightly underestimate the distance decay effect in the city. A possible explanation is still relevant to the habit of mobile phone usage: most people do not contact others by phone or text at every visited location. On average, displacements between phone calls (or text messages) are longer than those between consecutive locations people visit. Although it is true that numerous long-distance trips may be missing in the CDRs database as well, the amount of short-distance trips CDRs cannot "sense" could be substantially larger, which results in a less steep curve on the CDF plot for x < 20 km and a smaller value of decay parameter.

3.5.2 Community detection

Cell towers operated by MNOs can be considered as nodes in a huge cellular network. This network is capable of measuring human interaction with space. Like other types of network (e.g., social network), the network of a city often establishes a structure of communities, which are more tightly connected internally and

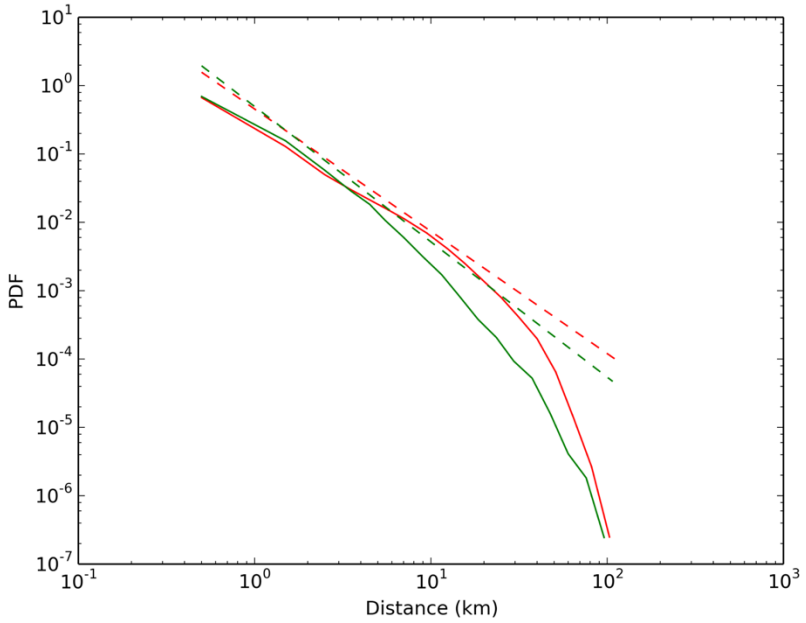

Figure 3.10 Probability density function (PDF) and the fitted power law distribution. The green line and red line represent the probability distribution of the displacements derived from the CDR group and the complete group, respectively. The dashed green line and the dashed red line are the fitted power law distributions for the CDR group and the complete group, with a decay parameter of 1.79 and 1.98, respectively.

structurally distinct from others (Girvan and Newman 2002). Identifying communities can help understand the internal structure of a city, shaped by human interaction with environment and urban infrastructure (e.g., land use, transportation), as opposed to pre-defined administrative boundaries. In recent years, some existing urban dynamics research detects urban communities using a vast amount of CDRs (e.g., Walsh and Pozdnoukhov 2011, Gao *et al.* 2013). Again, whether digital footprints that come with phone communications logs are biased for community detection needs to be examined due to the event-triggered nature of CDRs.

Community detection aims to partition a network into communities that consist of densely connected nodes. The quality of partition is often evaluated by modularity. In a weighted network, it is defined as:

$$Q = \frac{1}{2m}\sum_{i,j}[A_{ij} - \frac{k_i k_j}{2m}]\delta(c_i, c_j) \quad (3)$$

where A_{ij} denotes the weight of edge between two nodes i and j. $k_i = \sum_j A_{ij}$ denotes the sum of weight of all edges towards node i. c_i denotes the community node i is assigned to. $\delta(c_i, c_j)$ has a value of 1 if node i and node j belong to the

same community and a value of 0 if otherwise. m is half of the total edge weight in the entire network and $m = \frac{1}{2}\sum_{i,j} A_{ij}$ (Blondel *et al.* 2008). Numerous algorithms have been proposed to improve partition quality by maximizing modularity (e.g., Newman 2004, Clauset *et al.*).

Edges of a cellular network are usually weighted by the intensity of human interaction (i.e., volume of population flow). Given to the size of our network (33,044 cell towers), we adopt the Louvain method (Blondel *et al.* 2008), which takes a heuristic approach to optimize modularity of a large network efficiently. Using population movement volumes among cell towers in the entire day, the Louvain method is applied to detect urban communities using the CDR group and the complete group, respectively. As a result, we obtain optimistic community detection results given the high modularity scores (Table 3.9). For visualization purpose, we create a Voronoi diagram based on cell tower locations and assign a unique color to Voronoi cells in the same community.

Table 3.9 Summary of community detection results.

Group	No. of subscribers	Edges	No. of detected communities	Modularity
CDR	811,330	1,724,465	19	0.754
Complete	1,185,383	2,707,959	21	0.809

Figure 3.11 shows the 20 detected communities using data from the CDRs group. At the urban scale, the following findings are noteworthy:

1) Natural barriers play an important role in community separation. Two examples in Shanghai include the Yangtze River and the Huangpu River. The former separates the three islands of the Chongming County from other areas of Shanghai, while the latter divide Pudong and Puxi. As suggested by CDRs, although bridges and ferries provide means to transport people from one side to another, naturally separated regions remain sparsely connected in terms of the intensity of human interaction.
2) In many regions, administrative boundaries possess a similar power of separation as natural barriers do as the border of identified communities line up incredibly well with administrative boundaries. It implies that human movements within administrative districts are much more intense than cross-boundary movements. In other words, CDRs reveal that human interaction in Shanghai is largely affected by political boundaries.
3) Communities detected in Lujiazui area and Puxi, situated at the east bank and west bank of the Huangpu River, cover much smaller areas than others communities do. Compared with suburban districts (e.g., Jinshan, Songjiang, etc.), land use pattern in Lujiazui and Puxi, the most developed

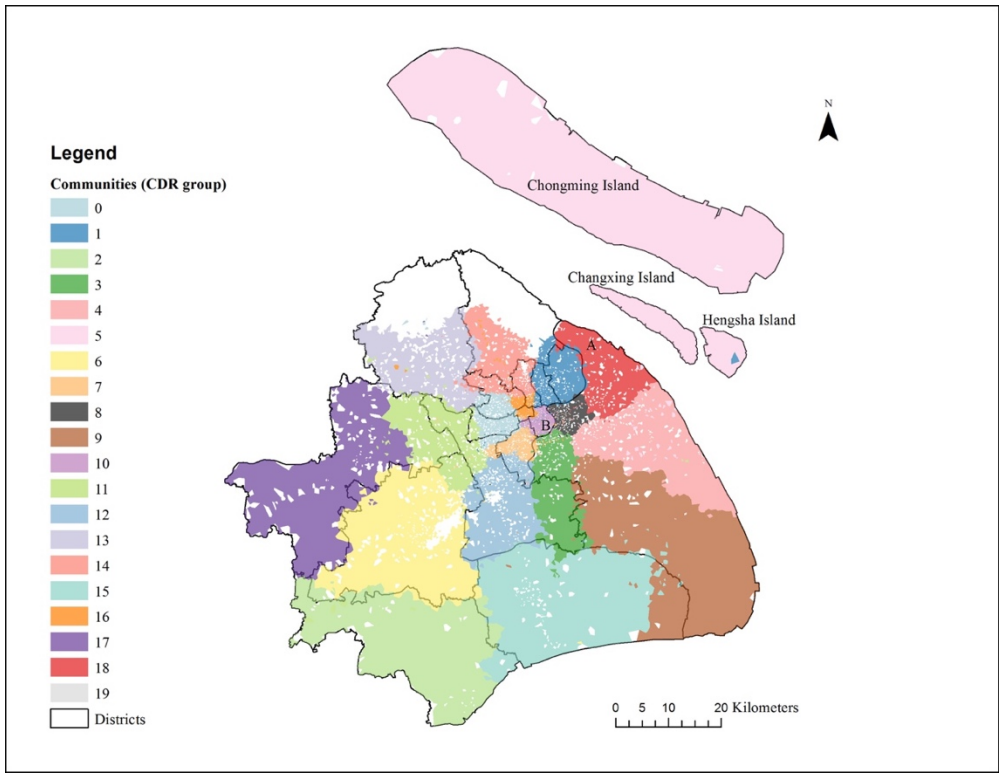

Figure 3.11 Detected communities based on the CDR group.

and most populated region in Shanghai, is highly mixed. Therefore, typical activities in this region do not require long-distance travel, resulting in smaller activity space on a workday.

By feeding data from the complete group, two more communities are identified. While the overall detection result well resembles the one from the CDR group in terms of the number of total communities and their boundaries, here we highlight and discuss some major differences:

1) In fact, natural barriers are not a decisive factor in community separation. For instance, Figure 3.12 shows that Region A, on the west bank of the Yangtze River, is closely connected to the three islands of the Chongming County (Chongming, Changxing, Hengsha, see Figure 3.12). Apparently, human interaction recorded in the complete group better capture population movement between Region A and the Changxing Island through ferries and a major arterial called the Changjiang Tunnel. Serving as critical parts of the Shanghai Port as a whole, industries related to port businesses (e.g., container ports, shipyards, shipping companies) are agglomerated in Region A and the Changxing Island. Region B, which covers both side of the Huangpu River in the south of the downtown area, presents another example in this case. These two regions are connected by the Nanpu

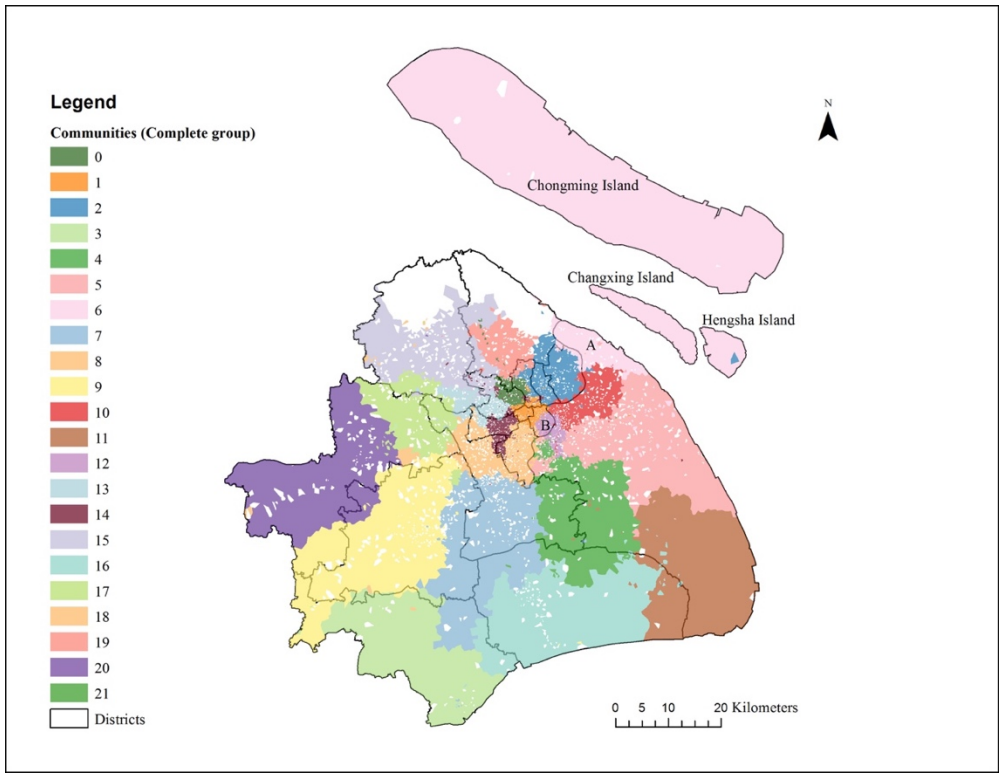

Figure 3.12 Detected communities based on the complete group.

Bridge, one of the main bridges over the Huangpu River. As a result, we learn that regions of a city might overcome natural barrier and become tightly connected communities. Unfortunately, we are not be able to uncover such relationship with CDRs because a large portion of people who travel between two naturally separated regions might not necessarily use their mobile devices in both regions.

2) Similar to natural barriers, political boundaries turn out to be less important in community separation. For example, with CDRs, communities in southern Shanghai are divided by administrative boundaries of Qingpu, Songjiang, Minhang, Jinshan, and Fengxian. With the complete group, the community in Minhang clearly crosses the administrative boundaries. Similar examples can be found in other places (e.g., the communities that cross Qingpu and Songjiang, Minhang and Pudong, Jiading and Baoshan, etc.). This finding suggests that although human interactions with space are more or less influenced by distinct socioeconomic environment of administrative districts (e.g., main industries), such influence is exaggerated by CDRs. Similar to the previous finding, we speculate the main reason to be the biased spatial distribution of calling/texting activities. For a large portion of subscribers, their primary phone communication activities might be limited at certain places, probably within the same administrative district.

3) By conducting a close visual inspection, we notice that in most cases, communities have clear boundaries and they are mutually exclusive: Voronoi cells that belong to one community seldom appear inside the territory of another. However, a number of exceptions are found at different locations of the city (see Figure 3.13 for an example), most of which cover residential area with high population density. This may be a result of home – work commute: a large percentage of residents travel to work outside the boundary of the community where their home is located, making the connection between the residential area and the destination community closer. On the contrary, communities detected using CDRs are more mutually exclusive: very few Voronoi cells are found to be inside a different community. This finding further confirms that CDRs can partially reflect intensity of human interaction in space. Tight connections across immediate neighborhood cannot be detected via phone communication logs.

Figure 3.13 Decreased level of mutual exclusiveness in community detection using data from the complete group. (a) Many Voronoi cells are located outside the community boundary. (b) Taking the Voronoi cell circled in the left figure as an example, we notice that most Voronoi cells outside the community boundary cover high-density residential area.

In this section, we adopt a collective approach to compare digital footprints from the CDR group and the complete group. Based on aggregated spatial

displacements, we investigate the distance decay effect and further realize how the limitation of CDRs could lead to biased understanding of urban dynamics. We also apply a community detection algorithm to identify the spatial structure of Shanghai in terms of human interaction. Although results from two datasets do not differ from each other by far, some major differences regarding how urban areas are separated by human interactions, such as those discussed in Section 3.5.2, cannot be ignored. Hence, we come up with a conclusion that from the collective perspective, urban dynamics patterns we uncover using CDRs might also be biased and aggregating data from individual level does not help address such bias.

3.6 Conclusions and Discussions

CDRs have been considered as an attractive data source in human mobility research. However, given the uneven distribution of people's communication activities in space and time, should we fully trust CDRs under all circumstances? This study takes the first step to understand the bias of CDRs in human mobility research. According to our evaluation, it is not realistic to simply answer this question with a yes or a no. First, it depends on what research question we aim to answer and the way we address it. For instance, CDRs tend to underestimate mobility indicators such as total travel distance and movement entropy. On the contrary, for certain problems that involve individual activity range, CDRs may be able to provide a decent estimate. Second, the effectiveness of CDRs is closely related to the habit of mobile phone usage. How frequently one uses mobile device to contact others, when and where those communications occur, largely determines the representativeness of his/her CDRs to the true mobility characteristics. In summary, we believe the event-triggered nature of CDRs does introduce certain degree of bias in human mobility research and we suggest people use caution in future research.

It should be noted that we do not attempt to deny the usefulness of CDRs in human mobility research. We may just have been excessively optimistic in the past. At present, CDRs remain one of the most useful data sources given large data volume and low cost in data collection and storage. Mining CDRs based on appropriate analysis techniques is still promising in academic research and real-world business. Thus, realizing potential problems of CDRs is more realistic than complete abandonment. Perhaps including some discussions of possible biases can be the first step. Next, it is worth thinking about possible ways to reduce/correct those biases if CDRs are the only available data source. For instance, interpolating CDR footprints can be one direction. It may help generate a more accurate estimate of certain mobility indicators, such as total travel distance. Applying post-hoc corrections is also a promising workaround. In Section 3.4 of this paper, we use linear regression to assess how much CDRs underestimate certain mobility indicators. The regression coefficient can be used to adjust

mobility indicator of interest. For instance, if we already know CDRs usually underestimate movement entropy of people who rarely use mobile phones by 50%, doubling the value of movement entropy derived from CDRs probably can yield a more accurate estimate. Nonetheless, our findings may be applicable in other cities due to different urban environment (e.g., socioeconomic status, transportation) and habits of mobile phone usage. A thorough understanding of local mobile phone usage patterns is necessary for post-hoc correction.

This paper only reveals the tip of the iceberg. Further research can be followed up to evaluate other approaches/methods we usually apply to analyze CDRs, in a more systematic way. A knowledge base that summarizes the effectiveness of CDRs under different scenarios can benefit future study. In the meantime, perhaps it is a good idea to rethink of existing findings the research community has come up with so far.

References

Ahas R, Silm S, Järv O, Saluveer E, and Tiru M 2013 Using mobile positioning data to model locations meaningful to users of mobile phones. *Journal of Urban Technology* 17: 3–27

Ascough II J, Maier H, Ravalico J, Strudley M 2008 Future research challenges for incorporation of uncertainty in environmental and ecological decision making. *Ecological Modeling* 219: 383–399

Batty M 2010 The pulse of the city. *Environment and Planning B: Planning and Design* 37: 575–577

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

Blondel V, Guillaume J L, Lambiotte R, Lefebvre E 2008 Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008: P10008

Brockmann D, Hufnagel L, and Geisel T 2006 The scaling law of human travel. *Nature* 439: 462–465

Brown L and Moore E 1970 The intra-urban migration process: A perspective. *Geografiska Annaler* 52: 1–13

Calabrese F, Pereira F, Lorenzo G, Liu L, and Ratti C 2010a The geography of taste: Analyzing cell-phone mobility and social events. *Lecture Note in Computer Science* 6030: 22–37

Calabrese F, Reades J, and Ratti C 2010b Eigenplaces: Segmenting space through digital signature. *IEEE Pervasive Computing* 9: 78–84

Clauset A, Newman M, and Moore C 2004 Finding community structure in very large networks. *Physical Review E* 69: 066111

Couclelis H 2003 The certainty of uncertainty: GIS and the limits of geographic knowledge. *Transactions in GIS* 7: 165–175

Delmelle E, Dony C, Casas I, Jia M, and Tang W 2014 Visualizing the impact of space-time uncertainties on dengue fever patterns. *International Journal of Geographic Information Science* 28: 1107–1127

Gao S, Liu Y, Wang Y, and Ma X 2013 Discovering spatial interaction communities from mobile phone data. *Transactions in GIS* 17: 463–481

Girardin F, Vaccari A, Gerber A, Biderman A, and Ratti C 2009 Quantifying urban attractiveness from the distribution and density of digital footprints. *International Journal of Spatial Data Infrastructures Research* 4: 175–200

Girvan M and Newman M 2002 Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America* 99: 7821–7826

Gollege R and Stimson R 1997 *Spatial Behavior: A Geographic Perspective*. New York, NY: The Guilford Press

Goodchild M and Gopal S 1989 *Accuracy of Spatial Databases*. London: Taylor and Francis

Zhang J and Goodchild M 2002 *Uncertainty in Geographic Information*. New York, NY: CRC Press

González M C, Hidalgo C A, and Barabási A-L 2008 Understanding individual human mobility patterns. *Nature* 453: 779–782

Griffith, D A, Millones M, Vincent M, Johnson D, and Hunt *A* 2007 Impacts of positional error on spatial regression analysis: A case study of address locations in Syracuse. *Transactions in GIS* 11: 655–679

Hägerstrand T 1970 What about people in regional science? *Papers of the Regional Science Association* 24: 6–21

Hecht B and Stephens M 2014 A tale of cities: Urban biases in volunteered geographic information. In *Proceedings of the International Workshop on Web and Social Media*. Ann Arbor, Michigan, USA: 197–205

Horton F and Reynolds D 1971 Effects of urban spatial structure on individual behavior. *Economic Geography* 47: 36–48

Jacquez G 2012 A research agenda: Does geocoding positional error matter in health GIS studies? *Spatial and Spatio-Temporal Epidemiology* 3: 7–16

Kang C, Gao S, Lin X, Xiao Y, Yuan Y, Liu Y, and Ma X 2010 Analyzing and geo-visualizing individual human mobility patterns using mobile call records. In *Proceedings of the 18th International Conference on Geoinformatics.* Beijing, China: 1–7

Kang C, Liu Y, Ma X, and Wu L 2012b Towards estimating urban population distributions from mobile call data. *Journal of Urban Technology* 19: 3–21

Liu Y, Kang C, Gao S, Xiao Y, and Tian Y 2012a Understanding intra-urban trip patterns from taxi trajectory data. *Journal of Geographical Systems* 14: 463–483

MacEachren A M, Robinson A, Hopper S, Gardner S, Murray R, Gahegan M, and Hetzler E 2005 Visualizing geospatial information uncertainty: What we know and what we need to know. *Cartography and Geographic Information Science* 32: 139–160

Mislove A, Lehmann S, Ahn Y, Onnela J, and Rosenquist J 2011 Understanding the demography of Twitter users. In *Fifth International AAAI Conference on Weblogs and Social Media*

National Bureau of Statistics in China 2012 Annual GDP of Major Cities in China. Available from: http://data.stats.gov.cn/workspace/index?m=csnd [Accessed 25 June 2015]

Newman M 2004 Fast algorithm for detecting community structure in networks. *Physical Review E* 69: 066133

Pang, A 2001 Visualizing uncertainty in geo-spatial data. In *Proceedings of the Workshop on the Intersections between Geospatial Information and Information Technology*, National Academies Committee of the Computer Science and Telecommunications Board, Washington, D.C

Perchoux C, Chaix B, Cummins S, and Kestens Y 2013 Conceptualization and measurement of environmental exposure in epidemiology: Accounting for activity space related to daily mobility. *Health & Place* 21: 86–93

Reades J, Calabrese, Sevtsuk A, and Ratti C 2007 Cellular census: Explorations in urban data collection. *IEEE Pervasive Computing* 6: 30–38

Refsgaard J C, van der Sluijs J P, Højberga A L, and Vanrolleghemc P A 2007 Uncertainty in the environmental modeling process – A framework and guidance. *Environmental Modeling & Software* 22: 1543–1556

Rhee I, Shin M, Hong S, Lee K, and Chong S 2011 On the levy-walk nature of human mobility. *IEEE/ACM Transactions on Networking (TON)* 19: 630–643

Rietveld P and Vickerman R 2004 Transport in regional science: The "death of distance" is premature. *Papers in Regional Science* 83: 229–248

Shanghai Municipal Statistics Bureau 2012 Land Area, Resident Population, and Population Density of Districts and Counties (2012). Available from: http://www.stats-sh.gov.cn/tjnj/nj13.htm?d1=2013tjnj/C0202.htm [Accessed 25 June 2015]

Sherman J E, Spencer J, Preisser J S, Gesler W M, and Arcury T A 2005 A suite of methods for representing activity space in healthcare accessibility study. *International Journal of Health Geographics* 4: 24

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

Song C, Koren T, Wang P, and Barabási A L 2010b Modeling the scaling properties of human mobility. *Nature Physics* 6: 818–823

Tobler W 2010 A computer movie simulating urban growth in the Detroit Region. *Economic Geography* 46: 234–240

Veregin H 1999 Data quality parameters. In Longley P A, Goodchild M F, Maguire D J, and Rhind D W (eds) *Geographic Information Systems*. New York: Wiley, 177–189

Walsh F and Pozdnoukhov A 2011 Spatial structure and dynamics of urban communities. In *Proceedings of the 2011 Workshop on Pervasive Urban Applications (PURBA)*. San Francisco, California, USA, 1–8

Wang Y, Lai P, and Sui D 2003 Mapping the Internet using GIS: The death of distance hypothesis revisited. *Journal of Geographical Systems* 5: 381–405

World Shipping Council 2013 Top 50 World Container Ports. Available from: http://www.worldshipping.org/about-the-industry/global-trade/top-50-world-container-ports [Accessed 25 June 2015]

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

Yuan Y and Raubal M 2012 Extracting dynamic urban mobility patterns from mobile phone data. *Lecture Note in Computer Science* 7478: 354–367

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

Zenk S N, Schulz A J, Matthews S A, Odoms-Yong A, Wilbur J, Wegrzyn L, Gibbs K, Braunschweig C, and Stokes C 2011 Activity space environment and

dietary and physical activity behaviors: A pilot study. *Health & Place* 17: 1150–1161

Zinszer K, Jauvin C, Verma A, Bedard L, Allard R, Schwartzman K, de Montigny L, Charland K, and Buckeridge D L 2010 Residential address errors in public health surveillance data: A description and analysis of the impact on geocoding. *Spatial and Spatio-Temporal Epidemiology* 1: 163–168

Chapter 4

Extract and compare generalized population movement patterns derived from different tracking datasets using a revised hierarchical clustering algorithm

Abstract

Understanding dynamic urban population movement patterns can benefit a variety of applications. In the past decade, researchers have been using tracking data collected by different means to better understand urban dynamics. Nowadays, lack of data is no longer the case in many developed regions. Instead, to answer a particular research question, we may have more than one tracking dataset and each of them reflects urban dynamics from a unique angle. This represents a new challenge in the big data era. In this paper, we aim to compare generalized urban population movement patterns of a big city in China, based on three tracking datasets: subway smartcard data, actively tracked mobile phone location data, and call detail record data. To effectively extract generalized population movement patterns, we propose a revised hierarchical clustering algorithm based on an existing publication. This revised algorithm groups OD flows in terms of their proximity distance and flow similarity. Results are discussed from three aspects: 1) we summarize urban population movement patterns revealed by three tracking datasets; 2) we demonstrate how each dataset differs from others and reveals urban population movement patterns from its unique perspective; 3) by combining conclusions from 1) and 2) and the characteristics of these datasets, we discuss their pros and cons in population movement analysis.

4.1 Introduction

Understanding dynamic urban population movement patterns can benefit a variety of applications, such as selecting locations for business and designing new routes for public transportation. Traditionally, researchers rely on travel surveys to study urban population movement (e.g., Crane and Crepeau 1998, Schlich and Axhausen 2003). In the past couple of decades, rapidly advancing information and communication technologies (ICT) have facilitated massive collection of digital footprints via various means, such as GPS-enabled devices (Zheng et al. 2008), social media (Batty 2010), at relatively low costs. Mobile phone location data have also drawn extensive attention in recent years due to the pervasive use of mobile phones (Ratti et al. 2006, Candia et al. 2008). Perhaps the most prominent feature of these datasets is the unprecedented scale. It offers the research community new opportunities to better understand population movement in a city.

Today, lack of data is no longer a problem in most developed regions. Instead, to answer a particular research question, we may have more than one tracking dataset and each of them reflects urban dynamics from a unique angle. In many cases, multiple tracking datasets collected in the same study area can tell different stories. As pointed out by Liu et al. (2015), such representativeness issue has become a top research priority in the big data era. It requires people to have a more thorough understanding of all available datasets in order to select the most appropriate one. This paper aims to extract and compare generalized urban population movement patterns of Shenzhen, a major city in southern China, based

on three tracking datasets: subway smartcard data, actively tracked mobile phone location data (we call it active phone tracking data in the remainder of this paper), and call detail record (CDR) data.

One fundamental challenge of visualizing and understanding movement data is the overlapping and cluttering issue (Andrienko and Andrienko 2010). Figure 4.1 shows 10% of the OD pairs derived from the active phone tracking data during 7-8 AM of a workday in Shenzhen. OD flows often overlap with each other to such an extent that one can hardly draw any meaningful conclusion. To effectively extract generalized population movement patterns, we propose a revised hierarchical clustering algorithm based on an existing publication (Guo 2009). This revised algorithm groups OD flows in terms of their proximity distance and flow similarity.

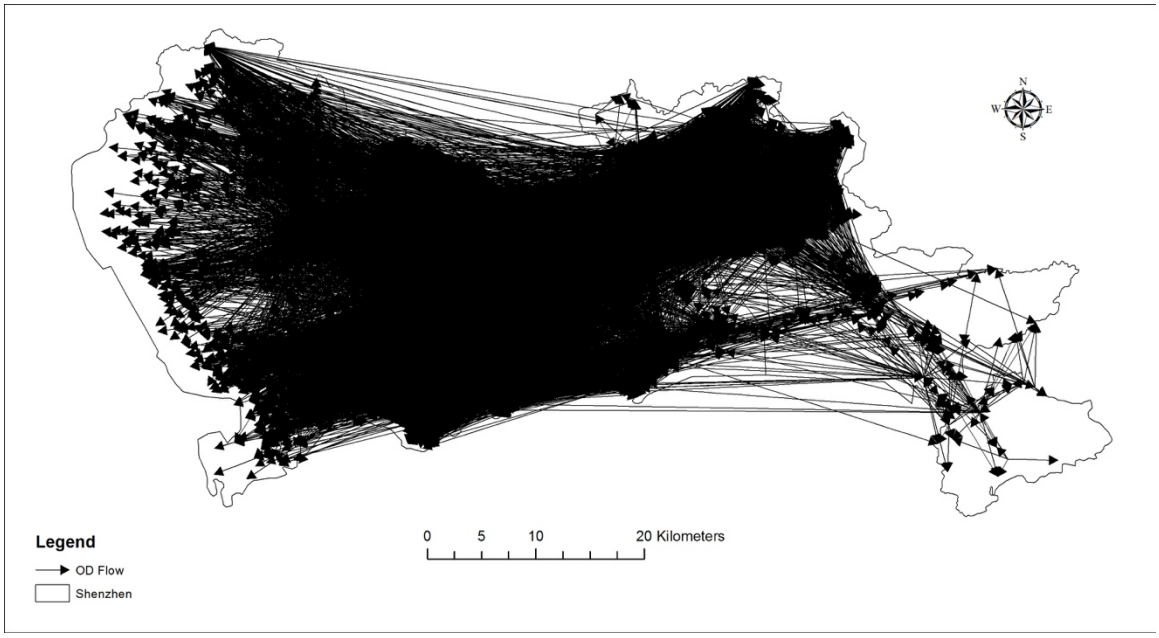

Figure 4.1 10% of the OD pairs among 5,952 cell towers during 7-8 AM of a workday in Shenzhen, China.

We then apply the proposed hierarchical clustering algorithm on each of the three datasets. Results are discussed from three aspects: 1) we summarize urban population movement patterns in two selected time periods, revealed by all three tracking datasets; 2) we illustrate similar and different population movement patterns uncovered from these three datasets; 3) we discuss the pros and cons of these three datasets based on their unique characteristics for population movement analysis.

4.2 Relevant research

This section discusses selected relevant research in the following two areas: 1) big tracking data and urban dynamics, and 2) flow data aggregation and clustering.

4.2.1 Big tracking data and urban dynamics

Rapid development of ICT has facilitated massive collection of digital footprints. It provides many opportunities in urban dynamics studies in terms of validating existing knowledge and discovering new insights.

Taxi trajectories collected by GPS have been considered as a valuable source for research of taxi drivers' behaviors and taxi trip patterns such as differences of "mobility intelligence" between top drivers and ordinary drivers (Liu et al. 2009a), distribution of trip direction and trip distance (Liu et al. 2012a), etc. Some studies focus on urban transportation environment reflected by taxi tracking data, such as distribution of pick-up and drop-off locations (Veloso et al. 2011) and "source-sink areas" (Liu et al. 2012b), identification of critical network locations (Fang et al. 2012, Zhou et al. 2015), and so forth.

Passenger trips recorded by public transit smartcards, which are originally designed as an alternative way of fare collection, also provide useful information of urban population movement (Pelletier et al. 2011). The majority of smartcard data research can be grouped into the following categories: 1) passenger behaviors such as number of transfers and trip durations (Park et al. 2008), variability of travel patterns (Morency et al. 2006), user group classification (Agard et al. 2006), and potential interactions (Sun et al. 2013); 2) current status of transit network (Liu et al. 2009b, Jang 2010); 3) analysis of dynamic urban system, such as job-housing relationships (Long and Thill 2015) and 4) future public transit planning (Utsunomiya et al. 2006).

Social media data are playing an increasingly important role in urban dynamics studies. Recorded digital footprints, such as location-based check-ins and geo-tagged tweets or photos, can be used to characterize individual or collective human mobility patterns (Preoţiuc-Pietro and Cohn 2013, Hasan et al. 2013, Azmandian et al. 2013). Many researchers consider users of social media as "sensors" in the socioeconomic environment (Goodchild 2007, Sagl et al. 2012, Liu et al. 2015). Besides monitoring the pulse of the city, those sensors are capable of distinguishing big social events (Lee and Sumiya 2010) or disasters (Sakaki et al. 2010).

In recent years, data collected from pervasive use of mobile devices has provided new insights in human mobility research (Ratti et al. 2006). Compared with other types of trajectory data discussed earlier, the scale of digital footprints in mobile phone location data is unprecedented. Most existing studies that analyze mobile phone location data are based on CDRs, which are originally used for billing

purposes. The main body of literature adopts an individual-based perspective and analyzes individual travel behavior based on CDR data (Gonzales et al. 2008, Song et al. 2010, Yuan et al. 2012, Gao et al. 2013, Xu et al. 2015). Others take a collective approach and focuses on aggregate mobility patterns in different urban areas (Girardin et al. 2009, Calabrese et al. 2011, Becker et al. 2013). However, one needs to be aware that CDRs are generated upon mobile phone usage. Each subscriber's location is recorded when he/she makes/receives a phone call or sends/receives a text message at the (x, y) coordinates of a nearby cell tower that handles the communication. In other words, a subscriber's locations are not recorded when his/her phone is not in use. Hence, depending on how actively one engages in phone communication activities, the number of recorded footprints of the same user in a day or among different users can vary drastically. Subscribers without any phone usage in a day are literally invisible in CDRs during that day. Furthermore, CDR data often reflect where people engage in phone communications rather than where they carry out various activities.

In the big data era, to address a specific problem, researchers and urban planners often face choices among multiple datasets. Given the unique properties of each dataset in terms of how digital footprints are generated, selecting appropriate data to answer different research questions has become a new and critical challenge in the big data era.

4.2.2 Flow data aggregation and clustering

Visualizing flows derived from migration data has been a challenging topic (Tobler 1976, Tobler 1987, Shaw et al. 2008, Guo 2009). In urban dynamics research, it becomes more difficult due to the vast amount of OD pairs in the city. Some widely adopted workarounds, such as filtering out minor flows (Tobler 1987), drawing OD flows with curved lines (Wheeler 2015), or sorting flows by volume and drawing them in an ascending order (Wood et al. 2011), can yield limited improvement. Reduction of dimension, which transforms OD flows to local in-flow/out-flow numbers, might be sufficient to uncover distinct mobility patterns at different urban locations (Guo et al. 2012). *Flowstrates* (Boyandin et al. 2011) also scarifies the spatial layout of OD flow by placing origins and destinations on two separate maps side by side. A heatmap view in the middle is used to illustrate varying flow volume on a particular OD pair. To preserve spatial layout, a useful approach is flow aggregation (Andrienko and Andrienko 2010). For instance, Andrienko and Andrienko (2011) develop a method that extracts key points from trajectories and groups them to a smaller set of locations using a point-based clustering algorithm. After that, original flows can be transformed into movements among those locations. A major shortcoming of this approach is that we lose median and long distance flows since they are all split to shorter ones. Inspired by graph theories, some researchers develop edge bundling processes, which merge nearby edges (flows) together (Cui et al. 2008, Holten and van Wijk 2009).

Different from point aggregation and edge bundling, Zhu and Guo (2014) develop a hierarchical clustering approach to cluster taxi flows in Shenzhen. This approach merges OD flows in terms of shared origins and destinations and progressively reduces the number of existing flows. It consists of the following steps:

1) Find k-nearest neighbors of each origin and each destination.
2) Identify neighboring flows $\{q_i\}$, for each flow p. O_p, D_p, O_q and D_q are used to denote the origin and the destination of p, and the origin and the destination of q. In addition, $KNN(O_p, k)$, $KNN(D_p, k)$, $KNN(O_q, k)$, and $KNN(D_q, k)$ denote the k-nearest neighbors of O_p, D_p, O_q, and D_q, respectively. For each q, to be qualified as a neighboring flow of p, the intersection of $KNN(O_q, k)$ and $KNN(O_p, k)$, and the intersection of $KNN(D_q, k)$ and $KNN(D_p, k)$ should not be null.
3) Calculate the proximity distance between each pair of neighboring flows using the following equation:

$$dist_p(p, q) = 1 - \frac{\left|KNN(O_p,k) \cap KNN(O_q,k)\right|}{k} \times \frac{\left|KNN(D_p,k) \cap KNN(D_q,k)\right|}{k} \qquad (1)$$

where $\left|KNN(O_p, k) \cap KNN(O_q, k)\right|$ and $\left|KNN(D_p, k) \cap KNN(D_q, k)\right|$ denote the number of shared nodes between $KNN(O_p, k)$ and $KNN(O_q, k)$, and the number of shared nodes between $KNN(D_p, k)$ and $KNN(D_q, k)$. If p and q share exactly the same neighbors, $dist_p(p, q)$ = 0. On the other hand, if no common neighbor is identified for either origins or destinations, $dist_p(p, q)$ = 1. Otherwise, $dist_p(p, q) \in (0,1)$.
4) Sort all n pairs neighboring flows in an ascending order by their proximity distance. The result of this step is an ordered list of neighboring flows.
5) Mark each OD flow as an independent cluster. This step creates n initial clusters.
6) Cycle through each neighboring flows, in the list generated in Step 4. If two flows are already in the same cluster, move on and process the next pair of neighboring flows. Otherwise, calculate the distance between the two clusters they belong to and merge the two clusters if their proximity distance is less than 1. The proximity distance between two clusters equals the proximity distance of the median flow of each cluster. If a cluster has only one flow, this flow is the median flow. Otherwise, the median flow is the one that is closest to the geometric center of the cluster.

This hierarchical clustering method is effective in revealing high-level population movement patterns from a vast amount of OD flows (e.g., taxi OD flows). Each OD flow participates in this process so information loss is controlled. However, it is not suitable for datasets used in this study. In the next section, we present the reasons and a revised hierarchical clustering method.

4.3 Method

Each OD flow in the taxi tracking data is unique (i.e., flow volume is 1) and both origin and destination are unique. However, this is very different from OD flows in this study, which are among a set of fixed facilities with varying sizes of population. The original hierarchical clustering method does not work in this case for the following two reasons: 1) the location of origins and destinations determines the clustering results and population size is completely ignored; and 2) distance between any two flows is determined by the proximity distance only and the characteristics of OD pairs are not considered. To address these issues, this paper proposes some revisions to the hierarchical clustering method such that it can better handle dynamic OD flows among a fixed set of facilities.

4.3.1 A revised hierarchical clustering algorithm

Figure 4.2 shows an example of six OD flows. It is evident that Flow 1 and Flow 2 are mainly movements during the morning rush hours, while Flow 3 and Flow 4 reflect movements mostly during the evening rush hours. On the contrary, population movements of Flow 5 and Flow 6 do not vary much in the day. These six OD flows suggest that varying population size throughout a day represents an important property of OD flows since it is closely related to aggregate spatiotemporal patterns of human dynamics.

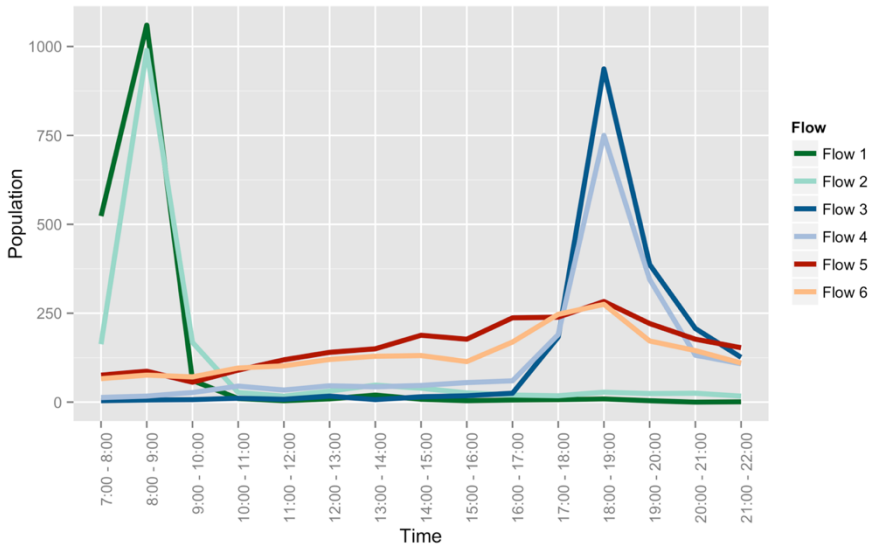

Figure 4.2 An example of six OD flows of population movements. Flow 1 and Flow 2 both peak during the morning rush hours, while Flow 3 and Flow 4 have their peaks during the afternoon rush hours. Flow 5 and Flow 6, on the other hand, remain relatively consistent throughout the day.

We take the varying population size into account and propose two major revisions to overcome the limitations of the original hierarchical clustering algorithm.

1) Distance measure:
With the revised clustering algorithm, proximity distance is calibrated by flow similarity, which represents how similar the population size of two OD flows varies over time. The calibrated distance between Flow p and Flow q is given by the following equation:

$$dist(p,q) = \begin{cases} 1 & if\ dist_p(p,q) = 1 \\ dist_p(p,q) \times dist_s(p,q) & otherwise \end{cases} \quad (2)$$

where $dist_p(p,q)$ is the proximity between p and q (see Equation 1) and $dist_s(p,q)$ is their flow similarity. The implication of Equation 2 is twofold:

a. The proximity distance is still one of the deciding factors. If two OD flows are not close to each other (i.e., they do not share k-nearest origins and k-nearest destinations), $dist(p,q)$ = 1. In other words, we do not adjust the distance between two OD flows if they are not geographically proximate.

b. If two OD flows are geographically proximate ($dist_p(p,q) < 1$), their distance is then adjusted by flow similarity, measured by their daily variations of population size. As discussed earlier, daily variation of population size reflects the dynamics of human mobility. OD flows with similar temporal population size variation patterns should receive a higher priority to be merged in the clustering process.

Flow similarity is calculated by the following equation:

$$dist_s(p,q) = \frac{\rho_{pq} \times (-1) + x}{x+1} \quad (3)$$

where ρ_{pq} is the Pearson correlation coefficient that ranges from -1 to 1 (i.e., from perfect negative correlation to perfect positive correlation). Since we aim to reduce the distance between flows with similar population size variation pattern, ρ_{pq} is multiplied by -1 in Equation 3. x is a factor that controls the influence that flow similarity has on the adjusted distance and $x \in (1, \infty)$. To make $dist(p,q)$ positive and meaningful, x needs to be greater than 1. Therefore, even if $\rho_{pq} = 1$, $dist_s(p,q)$ is still larger than 0. In general, a smaller x shortens the adjusted distance more significantly, whereas a larger x has a more limited influence. For example, if x = 3.0, $dist_s(p,q) \in [0.5, 1]$, which indicates that the distance between neighboring flows p and q can be shortened by 50% if their flow volume variation

patterns match perfectly (i.e., ρ_{pq} = 1). Hence, the distance between p and q is reduced by half so they can get merged at an earlier stage. In other words, they are less likely to be assigned to different clusters. If $x = 9.0$, $dist_s(p,q)$ ranges from 0.8 to 1.0, implying a less impact of flow similarity on adjusted distance. Theoretically, the value of x can be positive infinity. However, when x keeps increasing, $dist_s(p,q)$ approximates 1 and becomes less and less powerful in adjusting the proximity distance between p and q. Note that Equation 2 is the same as Equation 1 in the extreme case (i.e., $x \to \infty$).

2) Cluster merging criteria
After all neighboring flows are identified and the distances between them are calculated, individual flow clusters can be merged progressively to produce generalized flows. The original clustering algorithm looks for the median flow – the geometric center of the cluster – and uses this median flow to compute the distance between two clusters. Nevertheless, this approach does not take the population size into account. In reality, OD flows with higher population size (e.g., major arterials) play a more important role in a cluster than those with less population. In this study, instead of identifying the geometric center, the revised algorithm looks for the OD flow that is closest to the population-weighted center, and considers it as the center of the cluster.

4.3.2 Sensitivity analysis of k

Apparently, k is the most critical component in this hierarchical clustering algorithm. With a larger value of k, each OD flow is more likely to have more neighboring flows. This further implies a more thorough cluster merging process and thus a smaller number of final clusters.

In this section, we examine the influence of k by comparing the clustering outcomes under different settings of k, using 8,224 subway OD flows among 118 stations during 6-7 PM as an example. We run the revised hierarchical clustering by setting the k at 6, 8, and 10, respectively, and then examine the influence of k on the number of neighboring flows, neighbor search runtime, cluster merge runtime, and final number of clusters (Table 4.1). It indicates that when k is set to a larger value, the number of neighboring flows that can be identified increases considerably. As a result, the runtime for searching and merging neighboring flows becomes much more demanding. For instance, when k is set at 10, the neighbor search runtime and the cluster merge runtime grow to 86.65 seconds and 507.59 seconds. Compared with the runtime when k=6 (i.e., 31.40 seconds and 85.23 seconds), the computational intensity is escalated significantly. It is evident that the relationship between k and total runtime is not linear. Also, the final number of clusters decreases accordingly with a larger k.

Table 4.1 Influence of k on the number of neighboring flows, the neighbor search runtime, the cluster merge runtime, and the number of clusters.

k	Number of neighboring flows	Runtime: neighbor search (seconds)	Runtime: cluster merge (seconds)	Number of clusters
6	213,642	31.40	85.23	126
8	379,066	55.56	217.49	58
10	592,297	86.65	507.59	41

Figure 4.3 demonstrates the clustering results (k is set at 6, 8, and 10, respectively). In general, the revealed patterns are similar as indicated by major directions of population movements (see red circles). The main difference is the level of generalization. With k=6, although the population flow pattern revealed by 126 clusters (see Figure 4.3a) is already a high-level generalization of the original 8,224 OD flows, it is still somewhat challenging to differentiate smaller clusters due to slight overlapping issue. Setting k at 8 overcomes this problem since numerous smaller clusters are merged (Figure 4.3b). Figure 4.3c indicates that even more clusters are merged when k is set at 10. The population movement patterns become further generalized.

This section examines the influence of k in the revised hierarchical clustering algorithm. We learn that the setting of k has a large impact on computational intensity. From the perspective of derived population movement patterns, k controls the level of generalization effectively. It is worth noting that, although a large k leads to more generalized and understandable population movement patterns, we may risk missing local details as subtle clusters are merged to larger ones. It is wise to adjust k with different values in order to find the optimal balance between the level of generalization and the level of details.

4.3.3 Two-step hierarchical clustering

In the previous section we learn that hierarchical clustering is computationally intensive especially with an increasing k value. Computational intensity becomes a critical challenge when dealing with a network with thousands of nodes. For example, the active phone tracking data used in this study consist of 5,055,870 individuals moving among 726,119 OD pairs during 6-7 PM. When k is set at 10, the number of neighboring flows is 15,097,223 and the final number of clusters (81,641) remains very large. It takes almost one day to process this volume of OD flows with k=10, yet the level of generalization is still far from satisfactory. Although setting k at a larger value (e.g., 100) can help generate a more generalized and understandable result, it cannot be completed within a reasonable amount of time.

To overcome this computational issue (i.e., non-linear relationship between k and total runtime), we introduce a two-step clustering approach as a workaround. The

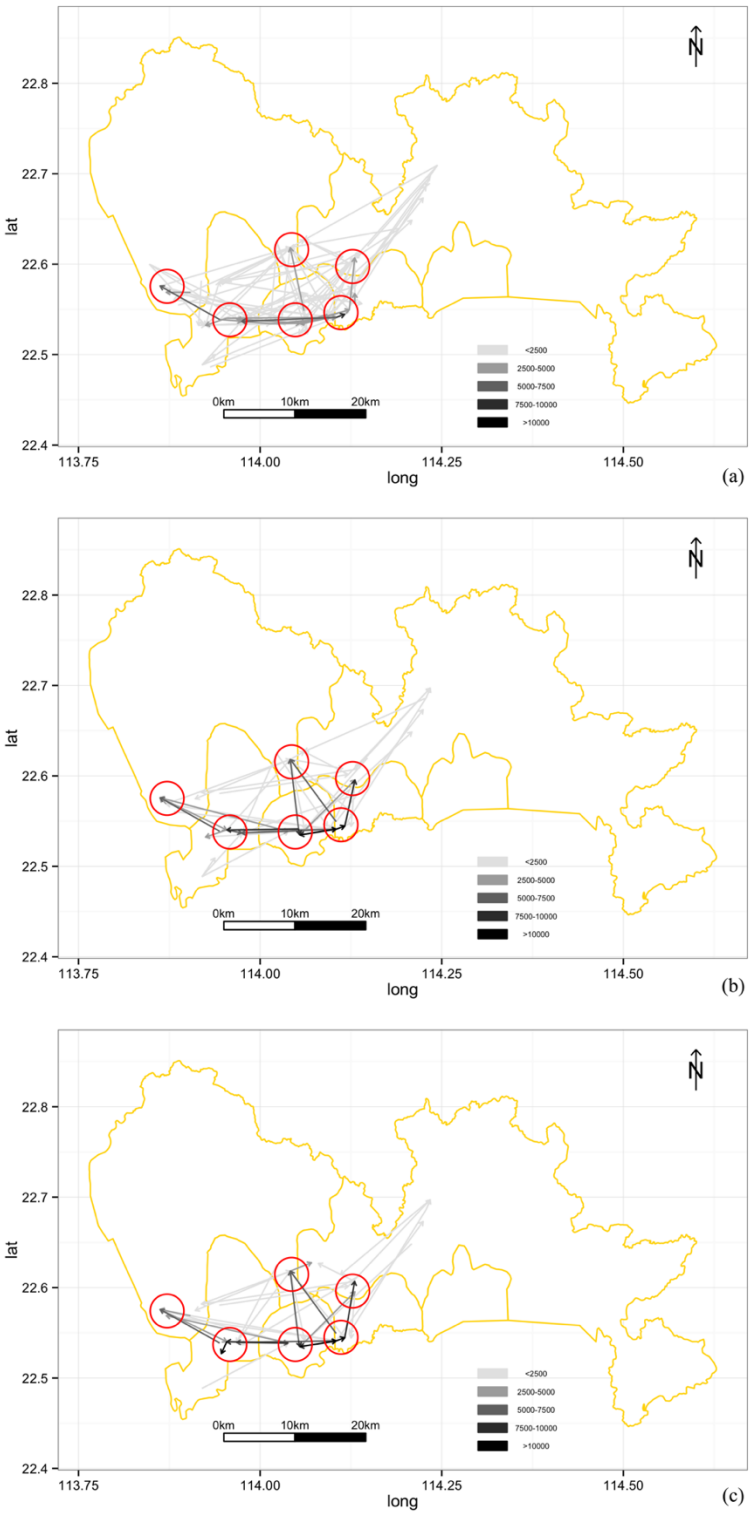

Figure 4.3 Subway flow clusters using different values of k. a) k=6 (126 clusters); b) k=8 (58 clusters); and c) k=10 (41 clusters).

concept of this approach is straightforward: we first use a small value of k to cluster the original OD flows with an acceptable computation time, and then apply a large value of k to produce a more generalized population movement pattern. The first step reduces the number of OD flows considerably and makes the second step – a more thorough clustering with a large k – feasible. Figure 4.4 displays the two-step clustering outcome for the subway OD flows during 6-7 PM. In this example, we use k=3 and k=8 respectively for the first step and the second step. The result indicates that, except for some slight differences, the population flow pattern revealed using the two-step clustering approach is very similar to that from the one-step clustering with k=10 (see Figure 4.3c). Major population clusters that emanate from the south to the north, and those move across the lower three districts are evident. In terms of efficiency, the total runtime reduces from 594.24 seconds to 24.09 seconds.

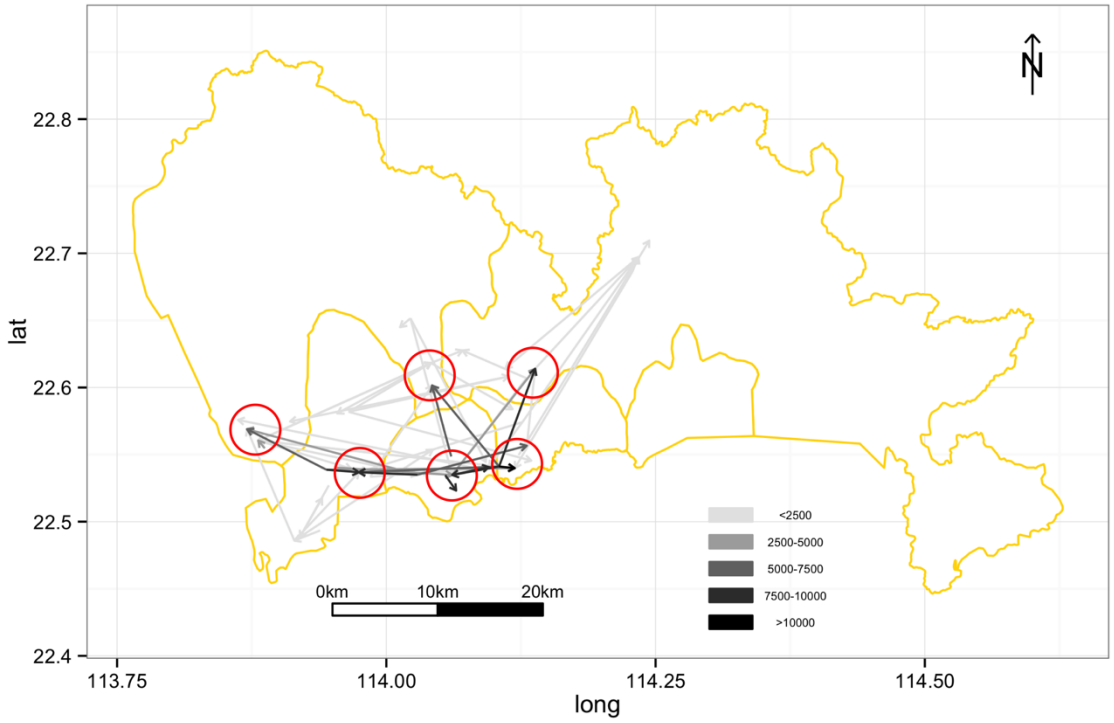

Figure 4.4 Subway flow clusters during 6-7 PM using the two-step hierarchical clustering (step 1: k=3; step 2: k=8; total number of clusters: 54).

The one-step hierarchical clustering approach is always preferred unless it becomes too computational expensive. In this case, the two-step flow clustering approach can help overcome the computational intensity issue. This approach is proved to be an effective workaround to group OD flows in a large network. In this study, we use this method to handle the active phone tracking data.

4.4 Compare generalized urban population movement patterns in Shenzhen

In this section, we apply the revised hierarchical clustering algorithm to extract and compare urban population movement patterns derived from the three different tracking datasets.

4.4.1 Study area and the datasets

The study area is Shenzhen, a major city in southern China located across the border from Hong Kong. After being designated as the first Special Economic Zone (SEZ) of China in 1980, its economy has been growing rapidly. Shenzhen is well known as a city of immigrants as it is the base of many manufacturing industries. Its population includes over 10 million permanent residents (see Gazette of the People's Government of Shenzhen Municipality 2012), most of who moved to Shenzhen during the last few decades, plus a substantial size of immigrants.

Shenzhen consists of six administrative districts (Figure 4.5). Longgang and Baoan, located in northern Shenzhen, are mainly for manufacturing industries. In the southeastern region of Shenzhen is Yantian, where a major port and many coastal resorts are located. The other three districts in southern Shenzhen (Nanshan, Futian, and Luohu) are known as the most economically developed parts of the city. Nanshan is a district designated to education and advanced technologies, while Futian and Luohu focus more on financial and other tertiary industries.

The subway smartcard data, the active phone tracking data, and the CDRs data are collected in October 2011, March 2012, and June 2011, respectively. All of them are generated on Friday. Although they are not the same Friday, we consider the overall urban dynamics do not change too much given the short interval.

As discussed in Section 4.2, the number of records associated with each subscriber varies significantly in a CDR dataset, depending on how actively a person engages in phone communications. To derive hourly population OD flows, we need to have a certain level of confidence that each trip occurs within a particular one-hour time window. For this reason, we process population OD flows using the following standard: assume a person leaves a footprint at cell tower c_1 at time t_1 and the next footprint is collected at cell tower c_2 at time t_2. If c_1 and c_2 are different and t_2 is within one hour from t_1, we consider this person moves from c_1 to c_2 during the one-hour time period. We exclude the trips of longer than one hour and focus on hourly population OD flows only in this study.

The active phone tracking dataset is similar to the CDR dataset in the sense that a subscriber's locations are recorded at the (x, y) coordinates of a nearby cell tower. However, individual footprints in the active phone tracking dataset are collected

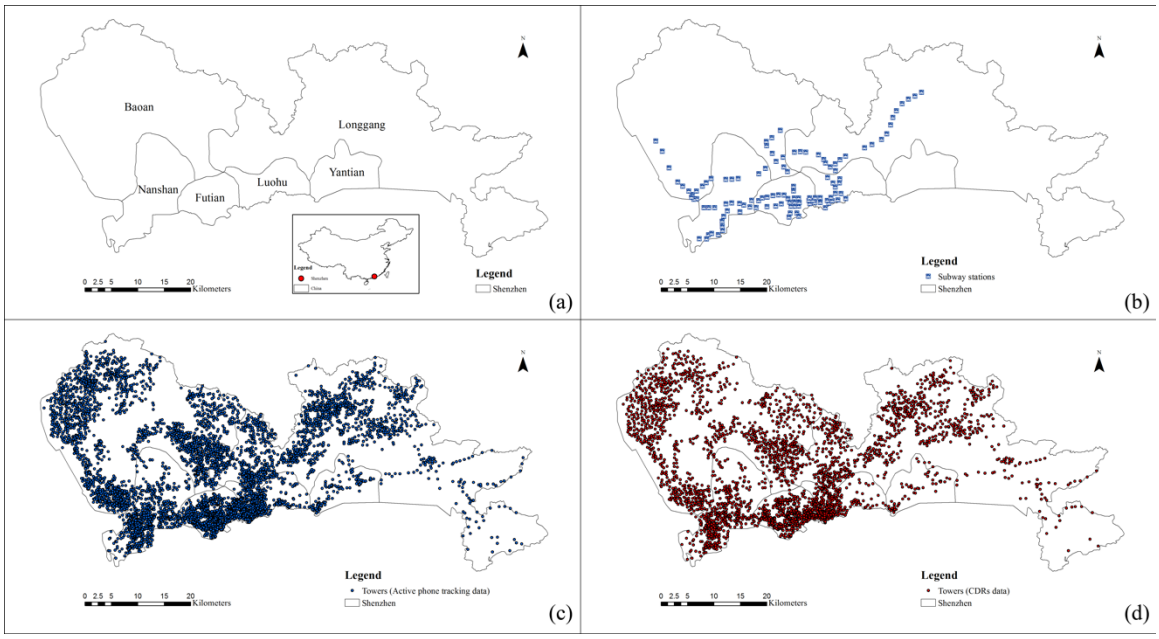

Figure 4.5 a) Shenzhen and its six administrative districts; b) Locations of all subway stations; c) Locations of all cell towers in the active phone tracking data; d) Locations of all cell towers in the CDR data.

every hour unless the mobile phone is powered off. This feature allows us to derive large-scale, hourly population OD flows in Shenzhen.

The smartcard dataset includes all subway trips made on a Friday among 118 subway stations. Each record contains a card ID, a timestamp, a ticket machine ID, and a transaction code, which indicates if a person enters or leaves a station. Aggregated one-hour population OD flows are computed using the same procedures we process the CDR data: a trip between station s_1 and s_2 is counted if the trip duration $(t_2 - t_1)$ is less than one hour and this trip is considered to occur in the one-hour time period that starts at t_1.

4.4.2 Results

Due to the space limit, we cannot present and discuss clustering results for each one-hour time period. Therefore, we select two representative ones: 1) 7–8 AM, a morning rush hour; and 2) 6–7 PM, an evening rush hour.

Table 4.2 lists some basic OD flow information regarding the total flow volume (i.e., population size) and the number of OD pairs. It is worth noting that the number of facilities (i.e., cell towers or subway stations) ranges from 118 to 5,952. This results in very large differences in the number of OD pairs. To produce flow clusters in a comparable degree of generalization, different values of k are necessary. We set k at 10 and 50 for the subway smartcard data and the CDR

Table 4.2 Summary of population OD flows of the three tracking datasets.

	Subway smartcard		CDRs		Active phone tracking	
No. of Facility	118		3,015		5,952	
Time Period	Volume	OD Pairs	Volume	OD Pairs	Volume	OD Pairs
7–8 AM	76,775	7,196	8,508	5,379	5,075,696	750,408
6–7 PM	141,502	8,224	80,810	13,363	5,055,870	726,119

data, respectively. Given the large number of OD pairs derived from the active phone tracking data, we apply the two-step flow clustering method by setting k at 10 in the first step and 50 in the second step in order to overcome the computational intensity issue.

We apply the revised hierarchical clustering algorithm for the two selected time periods of each dataset. The x parameter, which controls the influence that flow similarity has when adjusting their proximity distance (see Equation 3), is set as 3.0. We present and discuss clustering results in Sections 4.4.2.1 and 4.4.2.2.

4.4.2.1 Morning rush hour (7–8 AM)

Figure 4.6a demonstrates the subway flow clusters. During the time period of 7-8 AM, it is evident that most major flow clusters move towards the lower three districts in southern Shenzhen, including the base of many high-tech companies and research institutes in Nanshan and two major CBDs in Futian and in western Luohu. This illustrates the importance of these locations in attracting people from other regions and forming the distinct urban dynamics during the morning rush hour. In addition, there are some small clusters going from the south to the north, where many factories are located.

Based on the clustering results of the active phone tracking data, we can also observe evident flow clusters moving to Nanshan, Futian, and western Luohu. Nevertheless, the overall pattern is somewhat different from the subway flow clusters due to the characteristics of recorded movements in these datasets. Figure 4.6b indicates that all dominant flow clusters represent short-range movements, most of which are less than 2 km. These short-distance movements exist in a variety of urban areas, which suggests short-range home-to-work commutes do not concentrate in particular regions. Compared with results in Figure 4.6b, clusters derived from longer trips ($\geq 5km$) in the active phone tracking data illustrate the urban-scale population movement patterns more clearly since OD flows that are shorter than 5 km are removed (Figure 4.6c). Many large flow clusters move towards the lower three districts in the morning rush hour from different directions. In addition, it also suggests massive population movements in areas where existing subway services do not cover, such as a major expressway in western Shenzhen.

For the CDR data, the result indicates that all of them are very short local clusters

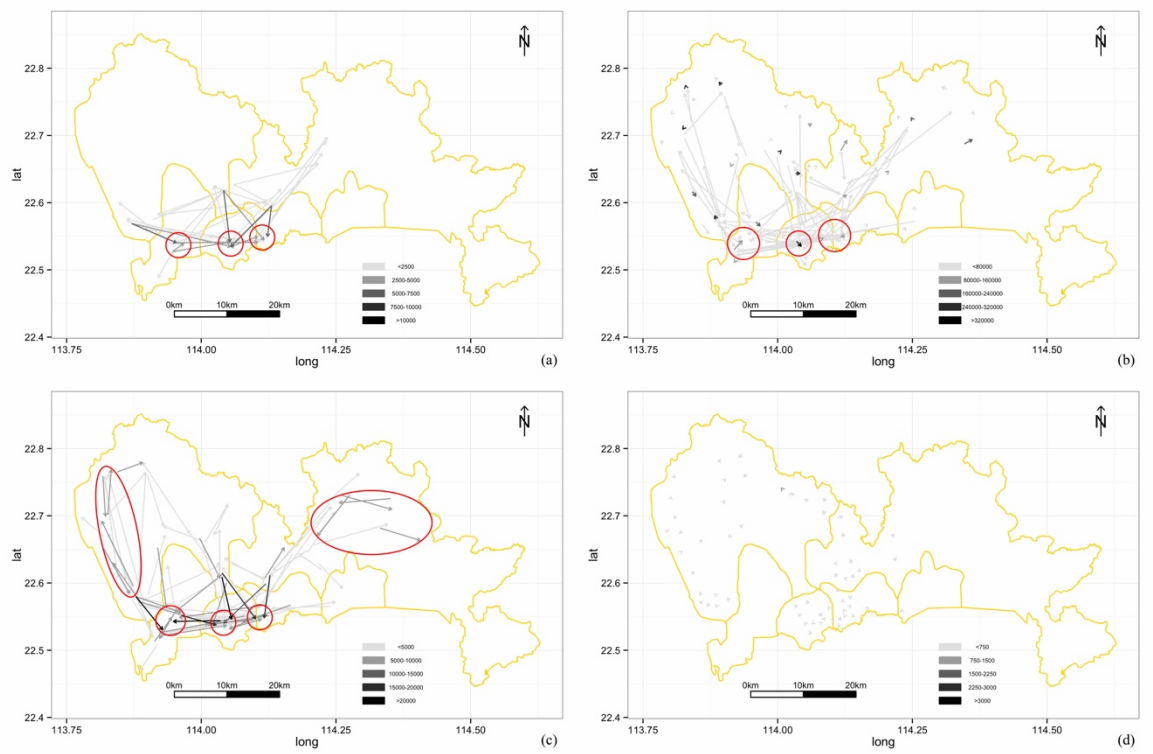

Figure 4.6 Hierarchical clustering results of the morning rush hour (7–8 AM) based on: a) subway smartcard data; b) active phone tracking data; c) active phone tracking data (trip distance ≥ $5km$); and d) CDR data. Note that 1) darker color represents a larger volume, and 2) to improve readability, very small OD flows which are not merged to main clusters are not displayed.

(Figure 4.6d). This makes sense because not too many people contact others during the morning rush hour (see Table 4.2). It is also interesting to observe that most people who use their mobile phones do not make long travels between phone communication activities.

In addition to visual inspection, we quantitatively evaluate movement directions of flow clusters in each administrative district. We define movement intensity of a direction as the product of the number of people (cluster size) heading that direction and their trip distance (length of cluster). The results are represented on a polar plot and directions are aggregated every 45 degrees (i.e., north, northeast, east, southeast, south, southwest, west, and northwest, see Figure 4.7). In general, for the subway smartcard data, prevailing movement directions in each district conform to the layout of tracks in that district. For instance, the majority of subway passengers in Longgang move towards the southwest direction to the lower three districts during 7–8 AM, while most Luohu passengers take westbound

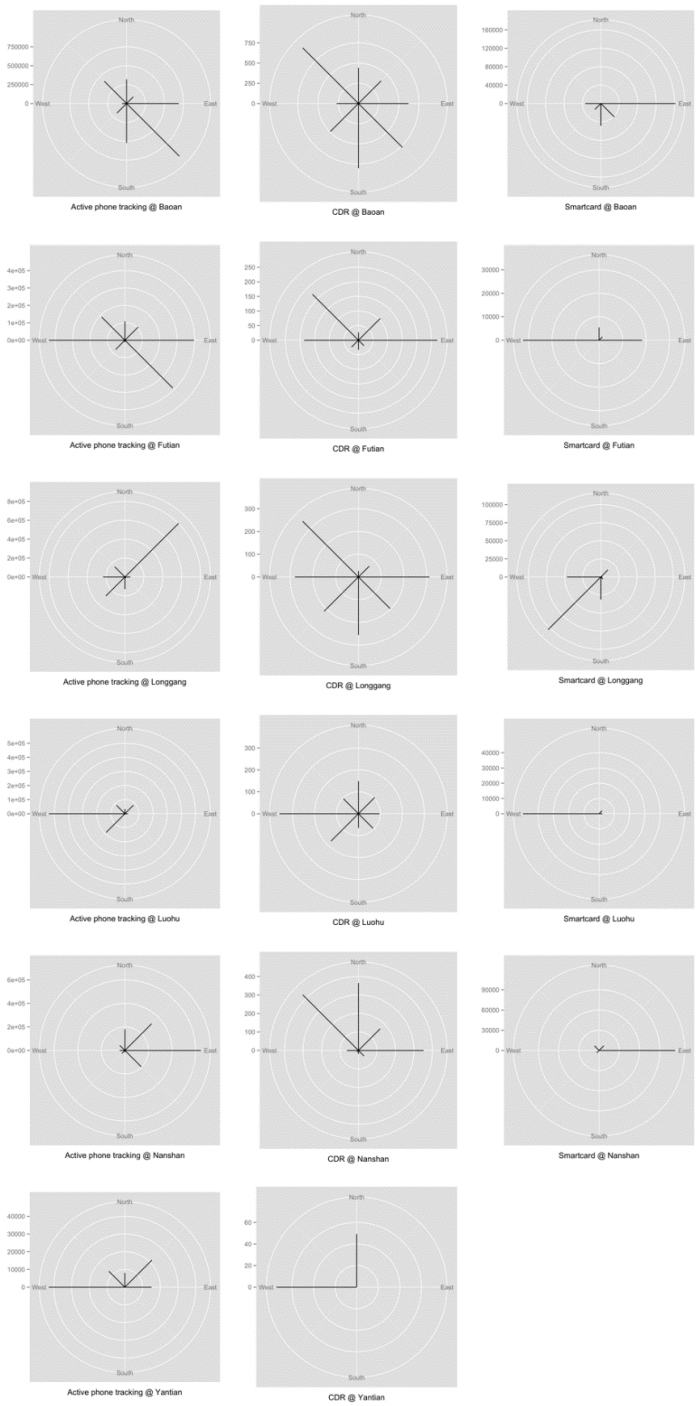

Figure 4.7 Flow direction distribution by administrative districts (7–8 AM).

trains to Futian and Nanshan. In the lower three districts where subway access is widely available, general movement directions suggested by the active phone tracking data do not deviate much from those derived from the subway smartcard data, whereas in Baoan and Longgang, clusters derived from these two datasets do not indicate similar trends of movement. Given the limited coverage of subway service, the subway smartcard data may underrepresent population movements in these two districts. The direction distribution of the CDR clusters appears to be more sporadic. Compared with the other two datasets, it does not reveal similar movement trend except the eastbound and westbound patterns in Futian.

4.2.2.2 The evening rush hour (6–7 PM)

Table 4.2 indicates that the subway system transports twice as many passengers during the evening rush hour of 6-7 PM than it does during the morning rush hour of 7-8 AM. The flow clustering results (Figure 4.8a) suggest that major flow clusters move in reversed directions (i.e., from the south to the north). This is a sign of work-to-home commutes. On the other hand, significant population movements across the lower three districts are also obvious. Our speculation is that a lot of people stay in this region after work for other activities, such as dinner and entertainment, due to the attractions in this region.

For the active phone tracking data, the population movement patterns in areas with subway service match well with the subway flow patterns, as suggested by the large number of south-to-north clusters and those across the southern region (Figure 4.8b). Clusters derived from longer trips ($\geq 5km$) in the active phone tracking dataset implies a more intense northward pattern, as we can see from the number of large flow clusters emanated from the lower three districts (Figure 4.8c). The result also indicates that many people use a major expressway in west Shenzhen.

For the CDR data, we do see more flow clusters in southwestern Shenzhen, which indicate people who use mobile phones and travel are mostly located in the more economically developed regions (Figure 4.8d). Again, this may be a sign of people staying in this region after work for other activities. However, it is difficult to uncover much useful information regarding urban-scale population movements due to the limitation of CDRs.

Distribution of movement direction further suggests some similarities and differences of three datasets (Figure 4.9). We notice that in many districts, the intensity values of the CDR clusters during 6–7 PM far exceed those in the morning rush hour, which comes from the fact that much more people engage in phone communication at 6–7 PM (except Yantian). Prevailing directions of CDR trips can be observed in Futian, Luohu, and Yantian, while this is not the case in other districts. According to subway flow clusters, certain direction(s) of movements are dominant in most districts. Major movement towards those directions are mostly

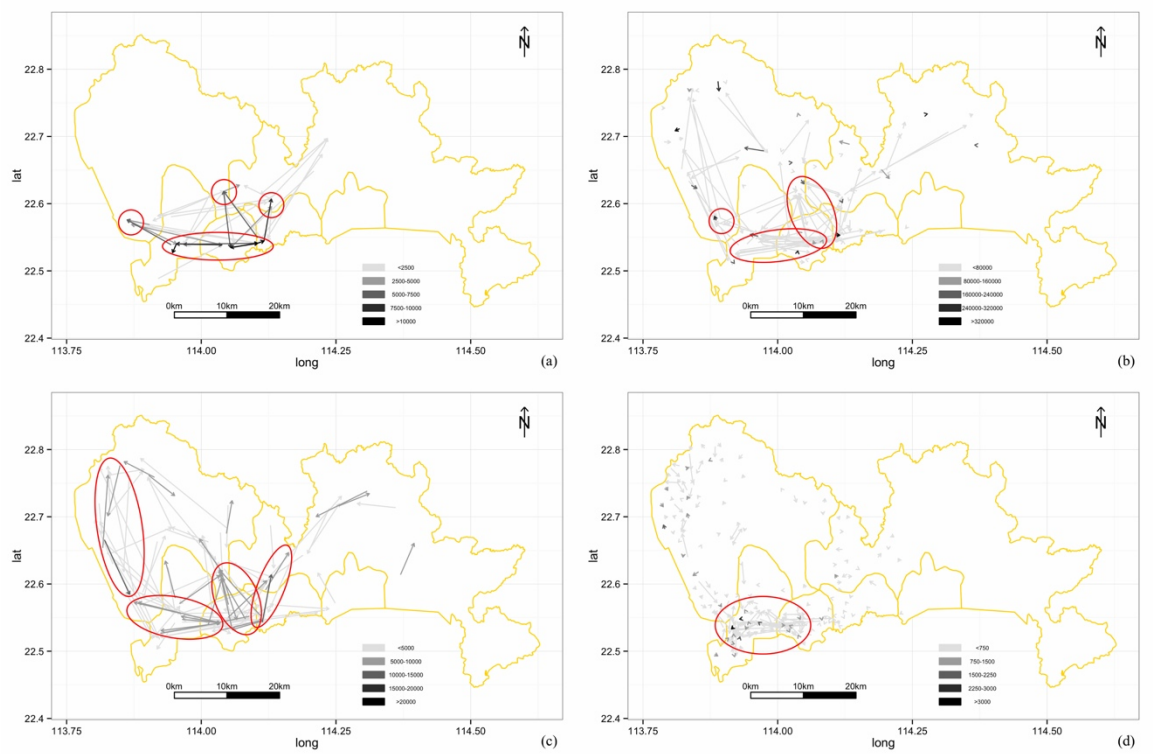

Figure 4.8 Hierarchical clustering results for the evening rush hour (6–7 PM) based on: a) subway smartcard data, b) active phone tracking data, c) active phone tracking data (trip distance ≥ $5km$), and d) CDR data.

supported by clusters derived from the active phone tracking data, while the latter also reveals other directions of movement that cannot be captured by the smartcard data.

4.4.3 Discussions

We believe that the clustering results are closely related to the characteristics of each dataset. The subway smartcard dataset records the origin and destination of each trip. Flow clusters based on this dataset clearly depict the distinct pulse of the Shenzhen subway system during the morning and the evening rush hours. In general, the active phone tracking dataset largely agrees with the subway smartcard dataset in terms of population movement patterns in areas with subway service. In addition, we find that longer movements are not evident since local clusters dominate urban dynamics in Shenzhen. For this reason, we also perform clustering analysis for trips longer than 5 km in order to better illustrate workday commute patterns. We learn that the results match subway flow patterns quite well and they improve our awareness of longer trips that occur outside of the subway system, such as massive population movements along major expressways.

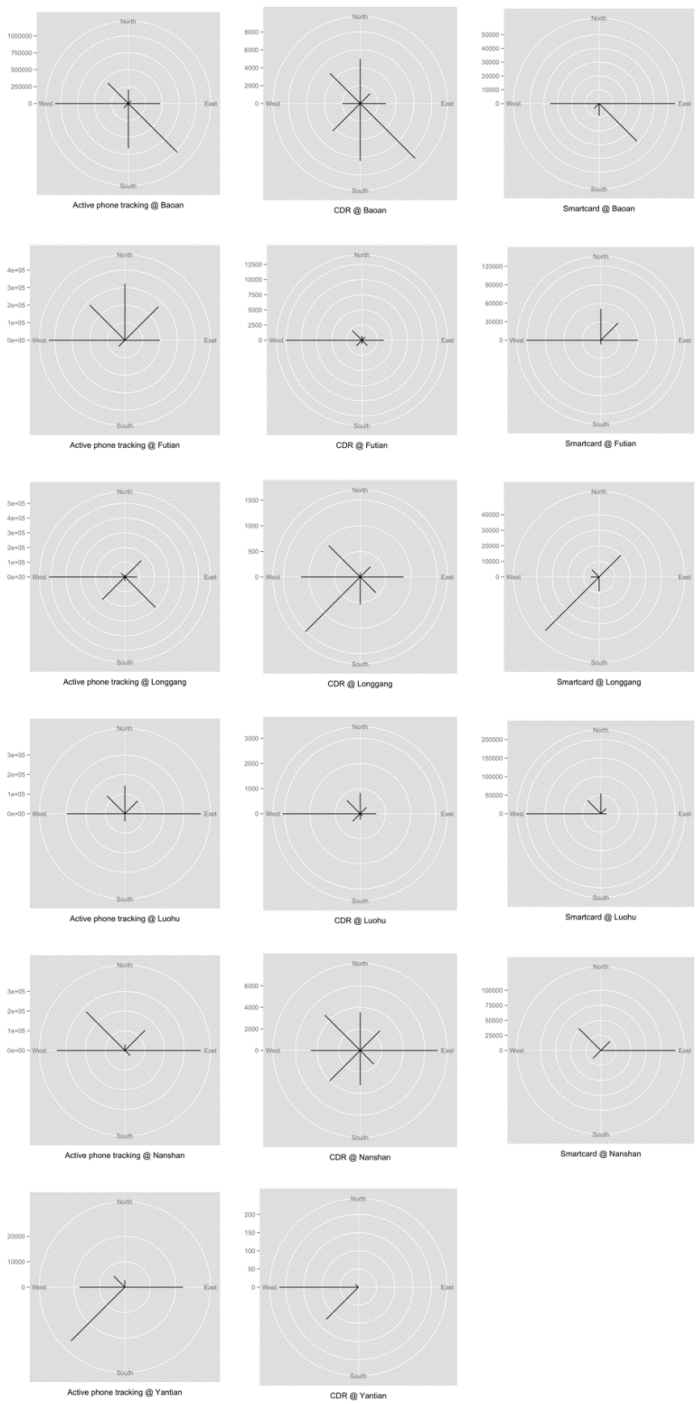

Figure 4.9 Flow direction distribution by administrative districts (6–7 PM).

The CDR dataset, on the other hand, is less useful in analyzing urban population movement patterns. Due to its event-triggered nature, it is only helpful in demonstrating the movement patterns of people when they use their mobile phone. During the morning rush hour of 7-8 AM, only 8,508 persons use their cell phone at two different locations (see Table 4.2). During the evening rush hour of 6-7 PM this number grows to 80,810. Such a big difference could be a reflection of mobile phone use pattern that a small number of people use their mobile phone during the morning rush hour, while far more people use their mobile phone (e.g., organizing evening activities, chatting with friends or families) during the evening rush hour.

4.5 Conclusions and future work

Understanding dynamic urban population movement patterns has a profound implication for researchers, urban planners, and policy makers. Pervasive use of mobile phones, new payment methods (e.g., smartcard), and emerging wearable devices provide various means of tracking people's location continuously at a relatively low cost. However, given the size of those datasets, extracting generalized population flows from millions of OD pairs remains a challenge.

In this study, we aim to extract and compare generalized population flows from three tracking datasets: 1) smartcard data, 2) active phone tracking data, and 3) CDR data. Inspired from an existing hierarchical clustering approach, we propose a revised algorithm to extract generalized flow clusters among a set of fixed urban facilities (e.g., subway stations, cell towers). This algorithm takes into account characteristics of each OD pair in terms of the varying size of population. As a result, when spatially proximate, similar OD flows (e.g., home-to-work commute trips) are more likely to be merged into the same cluster than dissimilar ones. Also, when merging clusters, the revised algorithm considers the population-weighted center of each cluster as the center of the cluster, instead of geometric center. Therefore, OD pairs with large population size have a greater influence on the distance between two clusters.

We apply the proposed hierarchical clustering algorithm to extract generalized population flows during two selected one-hour time periods for each of the three tracking datasets. The generalized population flows derived from the smartcard dataset reveal how people travel using Shenzhen subway system during the morning rush hour of 7-8 AM and the evening rush hour of 6-7 PM in a workday. We find that in the area where subway service is accessible, the overall pattern of generalized flows derived from the active phone tracking data largely agrees with the pattern derived from the smartcard data. The active phone tracking dataset offers benefits of identifying how people move in areas without subway access. The results can help urban transportation planners design new public transit services and optimize existing services. On the other hand, this study indicates

that CDR data are less capable of providing comprehensive urban population movement patterns due to its even-triggered nature. For example, uneven temporal mobile phone usage patterns could lead to a bias of digital footprints in CDR data. Therefore, we need to use caution when analyzing population movement patterns based on CDR data in urban dynamics research.

This paper serves as a starting point of urban dynamics studies using various types of tracking data. Continuous efforts can be made towards different future directions. On the algorithm part, several aspects of the proposed hierarchical clustering algorithm can be improved. For instance, currently we determine key parameters based on numerous experiments and our understanding of the data. To benefit the research community, a method that helps select key parameters (e.g., k and x) in terms of input tracking data is desirable. Also, the current implementation of the clustering algorithm is computationally intensive and it does not scale well. Its performance becomes a major limitation when the number of OD pairs exceeds certain level. A more efficient implementation that is capable of handling big OD flows will be very useful. For scientific research, it will be valuable to analyze population movement patterns based on multiple lengths of time window and examine how modifiable temporal unit problem (MTUP) affects our understanding of human mobility.

References

Agard B, Morency C, and Trépanier M 2006 Mining public transport user behaviour from smart card data. In *12th IFAC Symposium on Information Control Problem in Manufactoring (INCOM).* Saint-Etienne, France

Andrienko G and Andrienko N 2010 A general framework for using aggregation in visual exploration of movement data. *The Cartographic Journal* 47: 22–40

Andrienko N and Andrienko G 2011 Spatial generalization and aggregation of massive movement data. *IEEE Transactions on Visualization and Computer Graphics* 17: 205–219

Azmandian M, Singh K, Gelsey B, Chang Y-H, and Maheswaran R 2013 Following human mobility using tweets. *Lecture Notes in Computer Science* 7607: 139–149

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

Batty M 2010 The pulse of the city. *Environment and Planning B: Planning and Design* 37: 575–577

Boyandin I, Bertini E, Bak P, and Lalanne D 2011 Flowstrates: An approach for visual exploration of temporal origin-destination data. *Computer Graphics Forum* 30: 971–980

Calabrese F, Di Lorenzo G, Liu L, and Ratti C 2011 Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Computing* 10: 36–44

Candia J, González M C, Wang P, Schoenharl T, Madey G, and Barabási A L 2008 Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical* 41: 224015

Crane, R and Crepeau R 1998 Does neighborhood design influence travel?: A behavioral analysis of travel diary and GIS data. *Transportation Research Part D: Transport and Environment* 3: 225–238

Cui W, Zhou H, Qu H, Wong P C, and Li X 2008 Geometry-based edge clustering for graph visualization. *IEEE Transactions on Visualization and Computer Graphics* 14: 1277–1284

Fang, Z, Shaw, S-L, Tu, W, Li, Q and Li, Y 2012 Spatiotemporal analysis of critical transportation links based on time geographic concepts: a case study of critical bridges in Wuhan, China, *Journal of Transport Geography*, 23, 44–59

Gao S, Liu Y, Wang Y, and Ma X 2013 Discovering spatial interaction communities from mobile phone data. *Transactions in GIS* 17: 463–481

Gazette of the People's Government of Shenzhen Municipality. Issue No. 17, Serial No. 781, March 23, 2012. (available at http://english.sz.gov.cn/gg/201203/t20120329_1836874.htm)

Girardin F, Vaccari A, Gerber A, Biderman A, and Ratti C 2009 Quantifying urban attractiveness from the distribution and density of digital footprints. *International Journal of Spatial Data Infrastructures Research* 4: 175–200

González M C, Hidalgo C A, and Barabási A-L 2008 Understanding individual human mobility patterns. *Nature* 453: 779–782

Goodchild M 2007 Citizens as sensors: The world of volunteered geography. *GeoJournal* 69: 211–221

Guo D 2009 Flow mapping and multivariate visualization of large spatial interaction data. *IEEE Transactions on Visualization and Computer Graphics* 12: 1461–1474

Guo D, Zhu X, Jin H, Gao P, and Andris C 2012 Discovering spatial patterns in origin-destination mobility data. *Transactions in GIS* 16 411–429

Hasan S, Zhan X, and Ukkusuri S 2013 Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*. ACM, Chicago, Illinois, USA: 1–8

Holten D and van Wijk J J 2009 Force-directed edge bundling for graph visualization. *Computer Graphics Forum* 28: 983–990

Jang Wonjae 2010 Travel time and transfer analysis using transit smartcard data. *Journal of the Transportation Research Board* 2144: 142–149

Lee R and Sumiya K 2010 Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*. ACM, New York, NY, USA: 1–10

Liu L, Andris C, Biderman A, and Ratti C 2009a Uncovering taxi driver mobility intelligence through his trace. *IEEE Pervasive Computing* 160: 1–17

Liu L, Hou A, Biderman A, Ratti C, and Chen J 2009b Understanding individual and collective mobility patterns from smart card records: A case study in Shenzhen. In *Proceedings of12th International IEEE Conference on Intelligent Transportation Systems*. *St.Louis, MO, USA*: 1–6

Liu Y, Kang C, Gao S, Xiao Y, and Tian Y 2012a Understanding intra-urban trip patterns from taxi trajectory data. *Journal of Geographical Systems* 14: 463–483

Liu Y, Wang F, Xiao Y, and Gao S 2012b Urban land uses and traffic 'source-link areas': Evidence from GPS-enabled taxi data in Shanghai. *Landscape and Urban Planning* 106: 73–87

Liu Y, Liu X, Gao S, Gong L, Kang C, Zhi Y, Chi G, and Shi L 2015 Social sensing: A new approach to understand our socioeconomic environments. *Annals of the Association of American Geographers* 105: 512–530

Long Y and Thill J-C Combining smart card data and household travel survey to analyze job-housing relationships in Beijing. *Computers, Environment and Urban Systems (2015)*, http://dx.doi.org/10.1016/j.compenvurbsys.2015.02.005

Morency C, Trepanier M, and Agard B 2006 Analysing the variability of transit user behavior with smart card data. In *Proceedings of the 9th International IEEE Conference on Intelligent Transportation System*. Toronto, Canada: 44–49

Park J-H, Kim D-J, and Lim Y 2008 Use of smart card data to define public transit use in Seoul, South Korea. *Journal of the Transportation Research Board* 2063: 3–9

Pelletier M-P, Trépanierd M, and Morency C 2011 Smart card data use in public transit: A literature reivew. *Transportation Research Part C* 19: 557–568

Preoţiuc-Pietro D and Cohn T 2013 Mining user behaviours: A study of check-in patterns in location based social networks. In *Proceedings of the 5th Annual ACM Web Science Conference*. ACM, Paris, France: 306–315

Ratti C, Pulselli R M, Williams S, and Frenchman D 2006 Mobile landscapes: Using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design* 33: 727–748

Sagl G, Resch B, Hawelka B, and Beinat E 2012 From social sensor data to collective human behavior patterns – Analysing and visualizing spatio-temporal dynamics in urban environments. In Jekel T, Car A, Strobl J, and Griesebner G (eds) *GI-Forum 2012: Geovisualization, Society and Learning*. Wichmann Verlag, Berlin: 54–63

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

Schlich R and Axhausen K 2003 Habitual travel behaviour: Evidence from a six-week travel diary. *Transportation* 30: 13–36

Shaw S-L, Yu H, and Bombom L 2008 A space-time GIS approach to exploring large individual-based spatiotemporal datasets. *Transactions in GIS* 12: 425–441

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

Sun L, Axhausen K, Lee D-H, and Huang X 2013 Understanding metropolitan patterns of daily encounters. *Proceedings of the National Academy of Sciences* 110: 13774–13779

Tobler 1976 Spatial interaction patterns. *Journal of Environmental Systems* 6: 271–301

Tobler 1987 Experiments in migration mapping by computer. *American Cartographer* 14: 155–163

Utsunomiya M, Attanucci J, and Wilson N 2006 Potential uses of transit smart card registration and transaction data to improve transit planning. *Journal of the Transportation Research Board* 1971: 119–126

Veloso M, Phithakkitnukoon S, and Bento C 2011 Urban mobility study using taxi traces. In *Proceedings of the 2011 International Workshop on Trajectory Data Mining and Analysis*. ACM, Beijing, China: 23–30

Wheeler A 2015 Visualization techniques for journey to crime flow data. *Cartography and Geographic Information Science* 42: 149–161

Wood J, Slingsby A, and Dykes J 2011 Visualizing the dynamics of London's bicycle-hire scheme. *Cartographica* 46: 239–251

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

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

Zheng Y, Quannan L, Chen Yu, Xie X, and Ma W-Y 2008 Understanding mobility based on GPS data. In *Proceedings of 10th ACM Conference on Ubiquitous Computing (Ubicomp 2008)*. ACM, Seoul, Koreal: 312–321

Zhou Y, Fang Z, Thill J-C, Li Q, and Li Y 2015 Functionally critical locations in an urban transportation network: Identification and space-time analysis using taxi trajectories. *Computers, Environment, and Urban Systems* 52: 34–47

Zhu X and Guo D 2014 Mapping large spatial flow data with hierarchical clustering. *Transactions in GIS* 18: 421–435

Chapter 5

Conclusions

5.1 Summary

Understanding human mobility is essential for a diverse areas of real-world applications. The so-called “data-intensive paradigm” (Hey et al. 2009) and data-driven methodologies have dramatically enriched our knowledge of human activity and urban environment, as well as their interactions. Mobile phone location data, in particular, have helped reveal many novel aspects of human mobility, given the pervasive usage of mobile devices today. This dissertation research addresses three sets of questions proposed in Chapter 1 in three independent yet connected chapters.

Chapter 2 aims to uncover useful insights of urban dynamics based on a new data source: actively tracked mobile phone location data. By developing an innovative interaction visualization tool, some interesting patterns of urban dynamics in Shenzhen are disclosed, such as the dominant roles of stay activity and short-distance trip in human mobility. Inspired by these observations, the relationship between stay and move activities and the dynamics of the distance decay effect are then analyzed. The importance of stay activity is highlighted by calculating the ratio of stay population during each one-hour time period. Besides, this study takes the first step to evaluate the changing distance decay parameters in a day. In general, the frictional effect of distance varies significantly throughout the day and it is much weaker in the daytime due to the need for longer-distance travel. In addition, this chapter considers cell towers as sensors and proposes an approach to investigate the variation pattern of human mobility at different urban locations, measured by several mobility indicators derived from massive hourly digital footprints. Distinct spatiotemporal characteristics of human mobility across the city are revealed in terms of the changing volumes of stay, incoming, and outgoing population. To analyze different mobility patterns of the city, urban locations with similar mobility variation patterns are grouped together. Significant regional patterns, especially the major difference between the southern and northern parts of Shenzhen, are discovered. This chapter demonstrates the usefulness of actively tracked mobile phone location data in urban dynamics research.

Besides valuable insights that can be uncovered from the tremendous volume of digital footprint, the more consistent sampling rate of actively tracked mobile phone location data leads us to a more unbiased understanding of human activity. This remarkable feature also offers the possibility to validate our existing knowledge and re-think about previous CDR-based research. Based on a mobile phone location dataset that includes both CDRs and non-CDR footprints, Chapter 3 designs an innovative approach to assess the representativeness of CDRs. The entire dataset is separated into two groups: a CDR group and a complete group. From an individual perspective, selected mobility indicators that are frequently used in existing literatures are derived for each group. Disparities of mobility indicators between two groups are then evaluated and compared. From a collective perspective, this chapter analyzes the distance decay effect and detects

urban communities using digital footprints in each group. The results indicate that for individual mobility, CDRs are very likely to underestimate certain indicators significantly, whereas they might provide decent estimate for others. On the other hand, CDRs tend to provide biased understanding of collective human mobility as well. Another important conclusion is that the effectiveness of CDRs is closely related to the habit of mobile phone usage: how actively one uses mobile device to contact others, when and where those communications occur, largely determine the representativeness of his/her CDRs. This chapter takes the first step to carefully examine the bias of CDRs – the data source that has been widely used for years without ever being questioned.

Chapter 4 raises a new issue that researchers, urban planners, and decision makers have to consider: we now may possess more than one big tracking dataset to answer particular research questions, from urban planning to disease transmission. As a matter of fact, each type of data reflects urban dynamics from a unique angle. Thus, selecting the most appropriate dataset becomes so critical that it might affect the decision making process. Chapter 4 investigates this problem by extracting and comparing generalized population movement patterns derived from three datasets: subway smartcard data, actively tracked mobile phone location data, and CDR data. In order to effectively extract generalized movement patterns, this chapter develops a revised hierarchical clustering algorithm based on an existing publication (Zhu and Guo 2014). Compared with the original version, this revised algorithm takes into account the changing volume of moving population on each OD pair throughout the day and it groups OD flows based on two factors: proximity distance and flow similarity. The performance of the revised clustering algorithm is benchmarked using OD flow data at different scales. This algorithm is then applied to extract generalized population movement pattern from three different tracking datasets. Results indicate that in general, in areas where subway service is accessible, the overall pattern of generalized flows derived from the active phone tracking data largely agrees with the pattern derived from the smartcard data. On the other hand, CDR data are found to be less capable of providing valuable urban population movement patterns due to its event-triggered nature, which agrees with conclusions in the previous chapter. This chapter suggests researchers to use caution when analyzing population movement patterns based on CDRs.

5.2 Potential Applications

Although each chapter in this dissertation uses a particular dataset(s) collected in a particular area, some approaches can be applied under different contexts.

5.2.1 STEAM and geovisualization

In the big data era, geovisualization has become indispensable in urban dynamics research. It plays an important role in exploring data and setting up appropriate

research questions. Chapter 2 presents a program called *STEAM*, which is developed to visualize stay/move activities derived from actively tracked mobile phone location. To achieve a broader impact, *STEAM* has been released as an open source visualization package with a friendly graphical user interface (GUI) and a detailed user guide at *GitHub*: https://github.com/zlzhao1104/steam. In addition to the visualization capability demonstrated in Chapter 2, the GUI allows users to set a number of parameters, such as visualization speed, flow symbol, flow classification scheme, and so forth (Figure 5.1). Hence, urban planners and researchers from a variety of fields are able to use animation-based visualization to explore different types of flow data (e.g., human/animal migration, vehicle trajectories). It is also worth mentioning that *STEAM* is cross-platform so it can run on any major platforms with Java Runtime.

5.2.2 Variation pattern of urban dynamics

Chapter 2 of this dissertation develops an approach to classify urban locations in terms of the variation pattern of aggregate human mobility (i.e., stay population, incoming population, and outgoing population). It is proved to be very effective in mobility change detection.

This method has the potential to be used in the real world, especially in facility location selection scenarios. It has great advantages over survey-based approach since the scale and spatiotemporal granularity that come with actively tracked mobile phone location data can provide a more precise and timely estimate of how fast each location is gaining or losing population. For instance, public transit planners can design new bus/subway route in areas with large incoming/outgoing population but very limited public transit service. Using the same information, public transportation operators can match service schedules to in-flow/out-flow dynamics so as to improve efficiency. On the business side, real estate developers can choose to build office buildings at locations with large stay population during daytime, whereas taxi companies may boost income by dispatching drivers to areas with a substantial growth of outgoing population.

5.2.3 Hierarchical flow clustering

A city is a highly dynamic system, which consist of various types of moving elements (Guo 2009), such as flows of people, flows of vehicle, flows of goods, even flows of information. OD flows are now recorded with an increasingly finer granularity. For example, in the urban area, the density of cell towers has been growing every year. Hence, an effective flow clustering algorithm becomes critical to understand the dynamic human-city interaction. Despite the focus of Chapter 5 being comparing population movement patterns derived from different tracking datasets, the revised hierarchical flow clustering method developed in this chapter can be used in many scenarios. For instance, during an urban disease outbreak, it is vital to find out areas that require immediate medical actions. After feeding population movement matrix derived from mobile phone location data, the

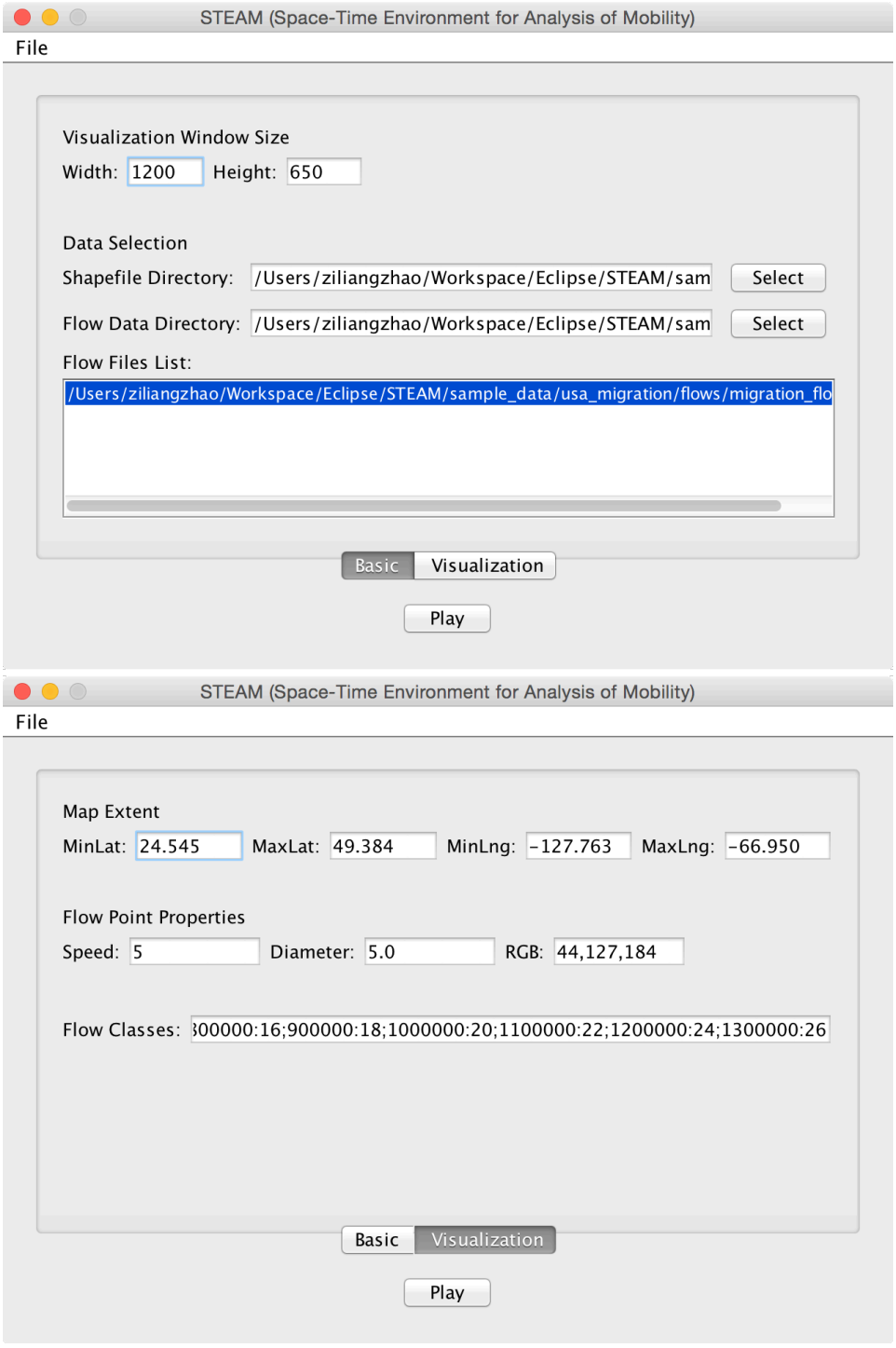

Figure 5.1 The graphic user interface (GUI) of *STEAM*. Users are able to set visualization window size and data directories in the "Basic" tab (the upper figure) and visualization-related parameters in the "Visualization" tab (the lower figure).

proposed flow clustering method can help identify major flow clusters emanated from the location of confirmed cases during the time period of initial outbreak. As a result, limited emergency response personnel and resources can be dispatched to those areas. Furthermore, after obtaining a good understanding of urban population movement patterns, the proposed flow clustering algorithm can be used to detect abnormal population movements caused by a wide range of events (e.g., political protest, natural disaster, social activity). Actions can be taken immediately to deal with the situation.

5.3 Future work

This dissertation research serves as a starting point of many potential directions of future work. Some of them are discussed in this section.

5.3.1 Validation of "patterns"

Many valuable findings of human mobility pattern derived from mobile phone location data are uncovered and elaborated in Chapters 3, 4, and 5. Nevertheless, those patterns are not validated because it is difficult to obtain relevant datasets. Unlike many western countries, many data sources (e.g., census data) are either unavailable or tightly controlled in China.

Further studies will be much more convincing if appropriate validation approaches can be developed. For instance, the distribution of home/work locations can be validated by population distribution data, while the generalized population movement patterns can be validated by some sort of travel survey. Alexander et al. (2015) provide a good example of validating conclusions inferred from mobile phone location data using census data and household travel survey.

5.3.2 The gap between "patterns" and "processes"

Like many other empirical studies, a major limitation of Chapter 2 is that the reason why a particular mobility pattern exists is discussed mainly based on speculations. *Google Earth* is the primary tool for determining major land use/building type covered by each cluster. For instance, high density residential areas are recognized in terms of the appearance of buildings in aerial photos. However, misjudgment can occur in this process. Consider this "residential building" example: in the real world, it is not uncommon that certain tall residential buildings are associated with extremely low occupancy rate. This is highly possible in China, where a considerable fraction of commercial housing is purchased for investment purpose. Although this chapter attempts to offer reasonable explanations for derived patterns based on a decent understanding of the study area, the reasons why a particular pattern is formed are still not certain. In other words, the underlying processes that drive human activities over space and time remain unclear.

The next step to continue Chapter 2 will be to bridge the gap between "patterns" and "processes". A prerequisite to make this happen is to obtain a number of supporting datasets so as to enrich actively tracked mobile phone location data, which is essentially massive OD flows at different time periods of a day. Land use data, population distribution data (e.g., census), POI data, will be very helpful not only because they can be used to explain particular patterns. More importantly, these datasets can be used to enrich digital footprints that mobile phone location data carry and make it more meaningful, such as "around 80% of trips between Location A and B during 7 AM and 8 AM are home->work commute trips", or "most people come to Location C for shopping". Jiang et al. (2013) discuss some initial efforts in enriching mobile phone location data, including activity/travel inference and matching digital footprints to road networks.

5.3.3 Data fusion and urban dynamics research

The so-called "big data era" is featured with three "V"s: volume, velocity, and variety (Laney 2001). Every day, an enormous volume of human-related data is generated from a variety of domains (e.g., social media, health care, transportation) with different representations, distributions, scales, and densities (Zheng 2015).

At present, the urban research community has not fully utilized the power of big data. Most existing studies of urban dynamics solely rely on data from one domain. Moving one step forward, Chapter 4 extracts and compares population movement patterns from three types of datasets and points out some pros and cons of each. Although this chapter highlights the necessity of selecting the most appropriate data for urban dynamics research, the promising opportunity lies in the integration of human-related data from multiple domains, which is also termed "data fusion". The benefits of data fusion ranges from improving data authenticity/availability, reducing data ambiguity, to enhancing data reliability, and so forth (Khaleghi et al. 2013). In urban dynamics research, since independent yet interconnected datasets collected in the same city reflect spatiotemporal characteristics of the urban system from different angles, an integrated approach of data analysis might yield a less biased and more accurate understanding of human activity. For example, smartcard records and taxi GPS traces can demonstrate "urban flows" with a fine spatiotemporal granularity. CDRs and check-in data from social media, on the contrary, are less capable of deriving the magnitudes and directions of human movement, whereas they can provide a good estimation of people's whereabouts.

Utilizing heterogeneous multi-source digital footprint is still rare in the urban research community. However, more and more researchers have realized the importance of this direction and started to develop methodologies which, according to Zheng (2015), can be classified into three categories: stage-based, feature level-based, and semantic meaning-based. With no doubt, data fusion will become

more prevalent and help uncover more practical insights for urban planners and stakeholders in the future.

This section discusses several potential research directions. The dissertation demonstrates the strengths of mobile phone location data in the new era of urban research. Meanwhile, this dissertation points out that it is equally important to study the limitations of mobile phone location data to ensure our enhanced understanding of human mobility and urban dynamics is correct, precise, and reliable.

References

Alexander L, Jiang S, Murga M, and González M C 2015 Origin-destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies* 58: 240–250

Batty M 2010 The pulse of the city. *Environment and Planning B: Planning and Design* 37: 575–577

Guo D 2009 Flow mapping and multivariate visualization of large spatial interaction data. *IEEE Transactions on Visualization and Computer Graphics* 12: 1461–1474

Gray J 2009 *The fourth paradigm: data-intensive scientific discovery*. Redmond, WA: Microsoft Research

Jiang S, Fiore G, Yang Y, Ferreira J, Frazzoli E, and González M C 2013 A review of urban computing for mobile phone traces: Current methods, challenges, and opportunities. In *Proceedings of the 2nd SIGKDD Workshop on Urban Computing*. ACM, Chicago, Illinois: 1–9

Zheng, Y., 2015. Methodologies for cross-domain data fusion: An overview. *IEEE Transactions on Big Data* PP (99): 1–18.

VITA

Ziliang Zhao was born in Fuzhou, China, to the parents of Xingyu Zhao and Xiaodan Lai. In 2005, Ziliang attended Jimei University in Xiamen, China, where he was introduced to Geographic Information Sciences. He obtained a Bachelor of Science degree from Jimei University in 2009. After that, Ziliang headed to the University of Tennessee to pursue graduate education. He graduated with a Master of Science degree in May 2011. Then, Ziliang continued his research at the University of Tennessee and obtained a PhD degree in December 2015.