# Understanding Activity Location Choice with Mobile Phone Data

Menglin Wang

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2014

Reading Committee:

Qiuzi Chen, Chair

Joshua E. Blumenstock

Jingtao Ma

Qing Shen

Program Authorized to Offer Degree:

Civil & Environmental Engineering

University of Washington

## Abstract

Understanding Activity Location Choice with Mobile Phone Data

Menglin Wang

Chair of the Supervisory Committee:

Qiuzi Chen, Associate Professor

Civil & Environmental Engineering

**Abstract.** Research on the variability in travel behavior, i.e. how individuals repeat or change their travel behavior over time, has a long history. An understanding of variability not only facilitates behavior modeling efforts but also provides insights into the complex factors underpinning people's travel behaviors. Previous studies of behavioral variability have mostly concentrated on non-spatial aspects. This dissertation contributes to the existing literature by examining the spatial variability of activity location choices. Activity locations are a set of spatially dispersed places where individuals perform activities. They play a critical role in structuring human travels, as individuals' demand for travelling in space is derived from the demand for activity participation. Specifically, two research questions are of interest: 1) how is location variability affected by time-of-day? 2) how can the knowledge of location variability be used to inform the development of location prediction models?

Analyses are performed with a mobile phone data set consisting of the traces of 120,435 individuals over two months. Significant time-of-day dependence of location variability is identified. Time-of-day effect is found to take account for 36% of the total variations in location

variability. The results emphasize the importance of time-of-day in shaping one's location choice behavior and provide a basis for future modeling efforts on location variability.

Location variability is found to be an instrumental indicator of the amount of input information required for location prediction. Specifically, given 100 historical (not necessarily unique) locations, an accuracy level marginally over 80% can be achieved for people with low location variability. In contrast, for those individuals with a high level of location variability, prediction accuracy can hardly reach 50% with 100 historical locations. This finding has significant implications on making more efficient location predictions. It will allow us to customize the amount of information input in location prediction for subpopulations differing in location variability by removing redundant information. Being able to discard some data without compromising model prediction accuracy is one way to deal with an overwhelming amount of data.

## TABLE OF CONTENTS

## LIST OF FIGURES

## LIST OF TABLES

## ACKNOWLEDGEMENTS

First, I would like to thank my advisor, Dr. Cynthia Chen. She led me out of the woods of graduate school to become an independent researcher. I'm also very lucky to have a very supportive dissertation committee. Dr. Jingtao Ma acquainted me with mobile phone data at the first place and introduced me to recent transportation applications of mobile phone data. Dr. Qing Shen has offered his expertise in urban planning to make sure this dissertation is embedded in a multidisciplinary context. Dr. James J. Anderson has been pushing me to think out of the box about my future research directions and spent a considerable amount of time to help me refine my writing. Dr. Joshua Blumenstock has been very helpful in broadening my research horizons by exposing me to his research with mobile phone data.

On a more personal note, there have been many friends I made during my graduate study. They have always been there for those ups and downs. Without their company, I probably would not have made this far on this road. Here, I would like to offer my sincere gratitude to Haiyun Lin, Manisa Veeravigrom, Fang Zong, Wilawan Thanatemaneerat and many others, who always stand by my side.

Finally and most importantly, I owe many thanks to my parents and my grandparents for the tremendous love and support they give me on this journey. It is them who make me who I am today. They have been, and will always be, the beacon by the sea helping me to find the way to the home in my heart.

# CHAPTER 1

# INTRODUCTION

## 1.1 RESEARCH QUESTIONS

Current data collection and modelling practices in travel behavior still largely rely on the assumption of highly repetitious behavior. There are reasons to expect repetition in travel behavior. First, human beings are regulated by biological clocks that prefer conducting certain activities on a regular basis (e.g. sleeping during night time every day). Second, human travel is subject to constraints (1) and recurring constraints necessitate repetitive behavior. For instance, morning commute is repeated at approximately the same time every day because the start time of work remains the same from one day to another. Third, people tend to avoid repetitive decision-making and thus repeat the same behavioral routines (2). Empirical studies also provide solid support for the existence of repetitive travel behavior (3, 4, 5).

While the existence of repetitious behavior is evident, variability in individuals' travel behaviors is also well recognized (6). The reasons to expect variability are as compelling as those for repetition. People are found to seek for variety in order to avoid the boredom of complete repetition (7). Also, unexpected circumstances can make routinized behavior impossible. Other reasons can result in variable travel behavior as well (e.g. learning process in adaption to a new environment) (2). Empirical results show considerable amount of variability in human travel behavior (2, 8, 9, 10).

Although the existence of variability in individuals' travel behavior is well acknowledged (2, 11), our knowledge of variability remains limited. Variability in travel behavior has received little attention in the literature, primarily because most data sets used for analyzing and modeling

urban travel comprise information for just a single day for each sampled individual and thus preclude examination of variability (12). This dissertation is devoted to the study of variability in travel behavior and its implications. Although studies have consistently shown that people exhibit distinct travel patterns between weekdays and weekends (13), this dissertation limits its scope to weekday travel only.

In this dissertation, variability in travel behavior is examined using mobile phone data. Mobile phone data sets can contain spatio-temporal positions of hundreds of thousands, even millions of mobile phone users, over an extended time period. It possesses enormous potential in behavioral research. While its application in behavior studies has seen great success, it remains to be explored by the travel behavior research community. This dissertation serves as one valuable addition to the existing literature on applying mobile phone data to travel behavior research.

Despite the spatial nature of human travel, previous research on variability in travel behavior has been largely focusing on non-spatial aspects of travel behavior. The amount of variability in travel behavior has been studied in terms of the number of trips, travel distance and travel time. These results do not necessarily represent the amount of variability in individuals' spatial behaviors. For instance, an individual can be observed to make the same number of trips on two days, though to completely different activity locations. Activity locations are a set of spatially dispersed places where individuals perform activities. They structure human travel in space in that individuals' demand for travelling in space is derived from the demand for activities participation (14). The primary purpose of this dissertation is to understand the variability in activity location choices, which is termed as location variability for short hereafter.

Location variability characterizes individuals' location choice behaviors and location choice has been shown to be affected by time of day (15). The first research question then arises is whether location variability depends on time of day. Different time periods in a day are filled with activities with different levels of fixity (16). Level of fixity is a measure reflecting the extent to which activities are constrained in space and time. Activities characterized with higher level of fixity are more difficult to be relocated and rescheduled. Therefore, participation in activities with a higher level of fixity tends to result in less variable activity location choices. Since daily activities vary in their level of fixity, time-of-day variations in location variability are anticipated.

In this dissertation, the dependence of location variability on time-of-day is analyzed by comparing individuals' location variability between different time periods of a day. In order to study possible population heterogeneity in this time-of-day effect, individuals are clustered into groups based on their levels of location variability. Lastly, the time-of-day dependence of location variability is quantified by introducing time periods as independent variables in explaining location variability.

One of the major reasons for understanding activity location choice is to develop more accurate and efficient location prediction models. Therefore, a second research question asks whether knowledge of location variability can be used to inform the development of location prediction models. Many location predictors require individuals' location history as input information. Longer history usually leads to more accurate predictions, but less computational efficiency. As model efficiency is becoming increasingly important for numerous applications relying on real-time location prediction, current attempts for improving prediction efficiency have been focusing on building more efficient predictors.

Little consideration has been given to reducing the length of input location history due to the concern of possible information loss. However, it can be shown that the amount of information carried by each observed location in a location history varies among individuals: for an individual who has lower location variability, repeated location choices contain little information that can be used to improve prediction accuracy. This observation suggests removing partial location history of those who are unlikely to vary their location choices would have negligible effects on prediction accuracy. Intuitively, for an individual who tends to repeat the same set of location choices every day, his location history on one day is sufficient to make an accurate prediction of his location choices on the next day and including additional location observations on previous days does not improve in prediction accuracy. In summary, it is hypothesized that location variability can serve as an indicator of the required amount of information in predicting individuals' activity location choices.

This hypothesis is tested by examining the correlation between history length and prediction accuracy. For subpopulations characterized with different levels of location variability, accuracy levels of a location prediction model are computed based on a set of input location histories with varying lengths. Variations in the levels of prediction accuracy with respect to the length of location history are then compared between subpopulations.

## 1.2 ORGANIZATION OF DISSERTATION

Examining location variability represents a dynamic perspective in studying human travel behavior. Most of the existing data collection and modeling practices in travel behavior domain are grounded in the conviction of highly repetitive travel behavior: few data sets used in

empirical studies have a length over one or two days and the majority of the models is static, i.e. models with no explicit time dimension. The assumption of routine travel behavior would be challenged if a considerable amount of variability in human travel behavior is identified. An insightful discussion of the significance of variability in travel behavior can be found in Jones and Clark (17).

Chapter 2 is included in this dissertation as a general background. This chapter is developed based on a paper recently submitted to *Transportation Research Part C*—Wang, M. and Chen, C., "Mobile Phone Data as an Alternative Data Source for Travel Behavior Research". In this chapter, I evaluate the potential of using mobile phone data as an alternative data source for travel behavior studies. Firstly, I briefly review current practices of using mobile phone data in travel behavior research. Secondly, I provide a description of mobile phone data sets in empirical studies, including their structure, information included and characteristics of the information, aimed at helping travel behavior researchers to develop an expectation for the type of information to be obtained from this type of data. Thirdly, I assess the advantages and limitations of mobile phone data within the context of travel behavior research to facilitate the evaluation of its appropriateness in specific applications. Lastly, I conclude that mobile phone data shows enormous potential for travel behavior research and remains to be exploited by the travel behavior community.

In Chapter 3, a detailed description of the data set used for analysis in this dissertation is provided. Various statistics are provided at different levels (i.e. individual level and population level), for a variety of measurements (e.g. number of sightings). These statistics provide us with valuable insights into the distinct nature of mobile phone data sets. Also detailed in this chapter is techniques used to process the data. Mobile phone data differs from conventional travel data in

multiple dimensions and, thus, requires non-traditional techniques[1] for data processing. The techniques described in this chapter can serve as a valuable reference for future efforts focusing on mobile phone data mining.

In Chapter 4, the first research question is answered—how time of day affects location variability. This chapter is formulated as a manuscript to be further developed as an journal article. Individuals are found to be more likely to vary their location choices in the afternoon than in the morning and evening. Yet, it is not the time-of-day dependence of location variability that I find surprising. Rather, it is how significantly time-of-day effect shapes individuals' location choice behaviors. Time-of-day takes account for approximately 36% of the total variations in location variability.

These findings emphasize the importance of time as a factor in influencing individual's travel and location choices. Even though its significance is well recognized, time has rarely been explicitly accounted for in location choice modelling (15). Only limited attempts exploiting the value of temporal information in location choice modelling (18, 19) can be identified. These studies have shown that location choice exhibits a prominent temporal pattern and characterizing location choice with time-of-day can potentially improve the power of location choice prediction. This part of my dissertation serves as a valuable addition to the research on the impacts of time on location choice behavior through quantified time-of-day effect on location variability and is expected to facilitate future efforts of location choice modeling and prediction.

Chapter 5 focuses on the second research question—application of knowledge of location variability for more efficient location prediction. This chapter also adopts a research paper

[1] Traditional techniques for travel behavior analysis are predominantly regression models.

format. Result show that levels of prediction accuracy differ for subpopulations differing in location variability, given the same length of location history. With only a dozen of historical locations, we can achieve an accuracy level marginally over 60% for people with low location variability. When the length of location history reaches 50 historical (not necessarily unique) locations, accuracy level climbs to over 70%. It keeps increasing to 80% as history length reaches 100 locations. In contrast, for those individuals with a high level of location variability, prediction accuracy starts at around 40% given ten prior locations and slowly increases to approximately 50% after 100 historical locations are observed.

These results have important implications on the development of more efficient location predictors. There is always a trade-off between prediction accuracy and model efficiency. More accurate prediction often requires more information. Yet, more information leads to higher processing cost. This trade-off becomes more salient in recent development of online location prediction applications emphasizing algorithm efficiency (20, 21). My results indicate that location variability can serve as an indicator for the amount of information used in location prediction models. Especially for individuals characterized with low location variability, location prediction model can perform well based on very limited history. Therefore, it is possible to improve prediction efficiency by removing a portion of input location history without compromising prediction accuracy. This will benefit many applications relying on both accurate and efficient location prediction.

# CHAPTER 2

# MOBILE PHONE DATA AS AN ALTERNATIVE DATA SOURCE FOR TRAVEL BEHAVIOR STUDIES

## 2.1 MOBILE PHONE AS A RESERCH INSTRUMENT

Mobile phones are becoming increasingly ubiquitous throughout the world. Recent market surveys show that mobile phone penetration rate has reached 100% in many industrialized countries (22). With mobile phone as sensors, studies have obtained novel insights into human mobility behavior. Mobile phone data has been explored in a variety of applications, including inferring social network structures (23, 24, 25, 26), understanding relationships between social interactions and physical locations (27, 28, 29, 30), monitoring population mobility (31, 32, 33, 34, 35), managing tourism (36, 37, 38, 39), and detecting social events (34, 40, 41).

In the transportation field, mobile phones have been used as probes for estimation of aggregate level traffic parameters, such as travel time (42, 43), travel speed (42), mode share (44, 45), origin-destination matrix (46, 47, 48, 49) and traffic volume (50, 51). Reviews of current practices using mobile phone as traffic probes can be found in (52, 53, 54, 55). On the other hand, mobile phone data also started to see its success in travel behavior research (56, 57). Travel behavior research is becoming increasingly important in recent years, as the focus of sustainable transportation is shifting from meeting travel demand by building more capacity to managing travel demand in order to maximize the use of the current transportation systems. Mobile phone data has allowed behavioral researchers to obtain valuable insights into travel behaviors from multiple levels, such as individual mobility patterns (56, 57), spatial interactions of socially connected people (23, 24), and population movements in large-scale space (58, 59).

Mobile phone data contains spatio-temporal locations of millions of mobile phone users over an extended period of time and provides an alternative way of colleting travel data. Before the advent of mobile phone data, travel diaries and Global Positioning System (GPS) tracking are two primary approaches of collecting disaggregate travel data and studying individual travel behavior. Both approaches are fraught with problems, among which prohibitive cost is the most commonly mentioned (60). The primary objective of this chapter is to introduce mobile phone data as an alternative data source for studying travel behavior. This objective is pursued from multiple dimensions.

First, studies on travel behavior with mobile phone data are reviewed and their most prominent behavioral findings are synthesized in order to inform travel behavior researchers with the most recent advancements. These studies, as empirical examples, illustrate the extent to which mobile phone data can facilitate behavior research and may potentially open up new avenues in travel behavior research. Moreover, these studies reviewed come from a diverse range of disciplines, including statistical physics (57, 61), sociology (27, 30) and computer science (62, 63) and thus present travel behavior researchers with new interdisciplinary research opportunities with the use of mobile phone data.

Secondly, a description of mobile phone data sets in empirical studies in provided, in terms of their structure, information included, characteristics of the information and current techniques for data cleaning. Mobile phone data is usually collected and owned by private-sector cellular network operators and researchers need to purchase the access the data. It is important for travel behavior researchers to develop an expectation for the type of information contained in mobile phone data sets before they take a serious interest in acquiring the data.

Lastly, the advantages and limitations of mobile phone data are accessed within the context of travel behavior research. It is important to note that collecting data is never an end unto itself, rather it serves as input into various applications. Therefore, the pros and cons of each type of data highly depend on the applications at hands. Though some advantages and limitations of mobile phone data have been recognized in previous studies (64), their implications on travel behavior study remain unclear. For this purpose, the characteristics of mobile phone data against travel data from alternative data sources are compared and an in-depth discussion on the potential issues with utilizing mobile phone data in travel behavior research is provided.

The rest of this chapter is organized as follows. In Section 2, a review of recent studies investigating travel behavior with mobile phone data is presented. Section 3 provides a detailed introduction to empirical mobile phone data sets, followed by Section 4 discussing both the advantages and the limitations of mobile phone data for travel behavior studies. The chapter is completed by conclusions on the prospects of mobile phone data in travel behavior studies in Section 5.

## 2.2 RECENT STUDIES WITH MOBILE PHONE DATA

In this section, recent developments in travel behavior research made with mobile phone data are reviewed. To facilitate our review, the existing literature is categorized into three groups based on the level at which travel behavior is investigated, namely microscopic study, mesoscopic study and macroscopic study. Studies employing a microscopic view focus on the characteristics of individual travel behavior; analyses carried at mesoscopic level make efforts to uncover the

interactions among the travel of a group of connected individuals; researchers adopting a macroscopic view are interested in aggregate travel of a population.

### 2.2.1 Microscopic: Characteristics of Individual Travel

Cellular network operators track the location of mobile devices to provide them voice and data services. Since people keep a phone near them most of the time, location of mobile device can be used to approximate individuals' trajectories in space and time. This line of research characterizes individuals' travel based on their trajectories reconstructed from location updates of mobile phone devices.

#### *2.2.1.1 Travel Distance*

Individuals differ in daily distance travelled. Studies (65, 66) show that the cumulative probability distribution of people's daily travel distance exhibited a skewed decay over larger travel distances. Specifically, while the majority of the sample covered a daily distance up to 10 km, some people regularly traveled as far as 100 to 300 km each day. Difference in travel distance seems to be related to city structure. People in Los Angeles were found to have a median daily travel distance two times greater than New Yorkers (67, 68).

Another commonly employed measure of travel distance is radius of gyration. Radius of gyration $r_g$[2] measures the size of an individual's trajectory and also exhibits population heterogeneity. In (57), the authors found that the distribution of $r_g$ followed a truncated power law: although most people's travel is confined to a limited area, some of us could regularly cover

[2] Let $L_i = \{l_1, l_2, \dots l_n\}$ be the sequence of visited locations of individual $i$ during the period of data collection. Then $r_g$ is defined by $r_g(i) = \sqrt{\frac{1}{n}\sum_{j=1}^{n} |l_j - \bar{l}|^2}$, where $\bar{l} = \frac{1}{n}\sum_j l_j$ is the center of mass of the trajectory.

an area up to hundreds of kilometers. This result was subsequently reproduced in (56, 57, 61, 65, 66) showing that the distribution of $r_g$ was fat-tailed.

*2.2.1.2 Regularity*

Individuals' travel is found to contain significant amount of spatial and temporal regularity from multiple dimensions. First, human beings tend to return to previously visited location (57). In (56), the authors investigated the number of locations visited for various windows of time. Though individuals tended to visit additional locations over time, a decrease in the rate of additional locations was evident and saturation was identified after three months. Song et al. (61) corroborated this finding by quantifying the relationship between the number of unique locations visited and time. They found the number of unique location visited followed a power function of time with a scaling factor smaller than 1, which means that the growth in the number of unique location slows down at large time scales.

Second, people devote most of their time to only a few locations and visit other locations with decreasing regularity. Gonzalez et al. (57) showed that, after each location visited by an individual was ranked by its visit frequency, the probability of finding this individual at a location was the inverse of its rank. Song et al. (61) reproduced this result by showing that the frequency of an individual visiting the $k$th most visited location $f_k$ could be approximated as $f_k \sim k^{-\zeta}$, where $\zeta \approx 1.2 \pm 0.1$. Furthermore, Song et al. (56) examined the fraction of time a user spent at his top-visited locations. The user was found to spend about 60% of his time at his top two locations. This number was reported to be 90% in a subsequent study (66). Lu et al. (65) elaborated these results by showing these percentages could differ depending on the number of unique locations visited. On average, those who visited more than 10 locations spent

approximately 75% of their time at the top two locations, while this percentage could be as high as 95% for those who only visited four distinct locations.

Third, human travel has a periodical nature. Gonzalez et al. (57) measured the return probability for each individual—the probability that a user returns to the position he/she was first observed after $t$ hours. The authors found this probability peaks at 24 h, 48h and 72 h, which highlights the daily rhythm of human mobility. Song et al. (56) aggregated the location visiting information for a sample over a course of six months and identified the most visited location for each of the 168 hours in a week. They observed that, on average, individuals would return to this most visited location during the same hour on the same day 70% of the time—a strong piece of evidence for weekly rhythm in human travel.

*2.2.1.3 Predictability*

Individual's travel has been proven to be highly predictable. Recent studies measured the predictability of human mobility with two related concepts: entropy $S$ and maximum predictability $\Pi^{max}$[3]. Entropy $S$ is a fundamental concept in measuring disorder in a time series and characterizes the degree of predictability. A user's trajectory with $S = 2$ can be interpreted as the uncertainty in this user's whereabouts is $2^S = 2^2 = 4$ locations. Maximum predictability $\Pi^{max}$ is the probability that an appropriate prediction algorithm can predict correctly a user's whereabouts. In other words, $\Pi^{max} = 0.2$ means we cannot predict a user's whereabouts with an accuracy level higher than 20%, no matter how good the prediction algorithm is.

[3] The relationship between $S$ and $\Pi$ is subject to Fano's inequality: if a user with entropy $S$ moves between $N$ locations, $\Pi^{max}$ is given by $S = H(\Pi^{max}) + (1 - \Pi^{max})log_2(N - 1)$ with $H(\Pi^{max}) = -\Pi^{max}log_2(\Pi^{max}) - (1 - \Pi^{max})log_2(1 - \Pi^{max})$.

Song et al. (56) showed, for 50,000 individuals, that their entropy $S$ peaked at 0.8 and correspondingly, the maximum predictability $\Pi^{max}$ peaked around 0.93, which means the uncertainty in a typical user's whereabouts is $2^{0.8} = 1.74$ locations and the predictability can be as high as 93%. Moreover, the distribution of both measures exhibited little variations across different subgroups defined by age, gender, home location and language. Under a different setting, Lu et al. (65) arrived at similar conclusions by showing the entropy of a typical user's was as low as 0.71, which leads to an uncertainty level as low as $2^{0.71} = 1.64$ locations. Not surprisingly, this low level of uncertainty resulted in a high level of predictability $\Pi^{max} = 0.88$. These results were proven to be rather robust even under extreme conditions. Lu et al. (66) investigated the mobility patterns of users affected by the earthquake in Haiti in 2010 and identified a slightly higher, even though still rather low, uncertainty level in human trajectories. They reported an uncertainty of $2^S = 2^{1.5} = 2.8$ locations and a maximum predictability $\Pi^{max} = 0.85$ for a typical user.

While maximum predictability sets the theoretical limit of prediction accuracy, a myriad of practical mobility prediction models have been developed in empirical studies and experimented with mobile phone data (19, 65, 69, 70, 71). These studies have reported a prediction accuracy level ranging from 60% to over 90%.

### 2.2.2 Macroscopic: Population Mobility

Location of individual mobile device, when aggregated, can be used to approximate the presence and flow of population in space. A significant number of studies have applied mobile phone data to study population mobility.

*2.2.2.1 Urban Dynamics*

Urban dynamics—research focusing on the temporal variations of population distribution in urban environment—has experienced significant developments aided by mobile phone data. Different experiments have been found in Rome (Italy) (72, 73), Milan (Italy) (59, 74), Shenzhen (China) (75) and Estonia (35). Mobile phone data has been proven to be a promising tool in monitoring population mobility in large urban space at different time scales. Population flow in urban environment was found to follow repetitious temporal patterns (e.g. daily pattern). Moreover, these temporal patterns possess surprising similarities across different urban environments: for four cities in Northern Italy, the total population in the city reached maximum on Tuesday and the minimal total population was observed on Sunday (76).

It is also possible to target the movements of subpopulations with mobile phone data. Ahas et al. (32) monitored the movements of suburban commuters' in the city of Tallinn, Estonia. Girardin et al (77) used mobile phone data to investigate the movements of visitors to a major exhibit in New York, 2008. Ahas et al. (37) analyzed the mobility of foreign tourists' in Estonia based on their mobile traces. Wesolowski and Eagle (78)(78)(78)(78)(78)(78)(78)(78) paid special attention to the mobility of slum dwellers in Kenya.

*2.2.2.2 Urban Structure*

A series of studies were able to produce valuable insights into the functional configuration of urban space with mobile phone data (79, 80). As an example, in a study conducted at Morristown, New Jersey (31), the authors applied mobile phone data to identify the set of residential areas that contribute most workers to a city. As an another example, Vieira et al. (81) proposed a new algorithm to identify dense areas in an urban environment, i.e. areas that are intensively used

within a time period, with mobile phone data. Moreover, multiple studies show that it was possible to identify home and work locations of city inhabitants from mobile phone data (64, 82, 83, 84).

*2.2.2.3 Special Events*

It is possible to use mobile phone data in detecting special events (75, 85), from social events to natural disasters. Specifically, Sagl et al. (76) used mobile phone data to detect soccer matches and Traag et al. (40) further showed examples with football games and musical festival. Mobile phone data is also a promising tool to learn human mobility during special events. Bengtsson et al. (86) validated the use of mobile phone data in determining population displacement after Haiti earthquake and monitoring population movements during a disease outbreak. Bagrow et al. (87) applied mobile phone data to study population mobility and interaction during eight real-world emergencies, such as bomb attacks.

### 2.2.3 Mesoscopic: Travel and Social Interaction

Since mobile phone serves as one of the major vehicles of communication, social interactions can be captured via calling activities. Therefore, mobile phone data has also intrigued research interest in the interplay between individuals' movements in the physical world and their interactions in cyberspace.

*2.2.3.1 Interplay between Mobility and Social Interaction*

Individuals' mobility is greatly shaped by the movements of others in his social network. Eagle et al. (24) investigated the temporal and spatial patterns of physical proximity of friends and found that friends were more likely to spend off-work hours at non-work places. In (30), the

authors found that 80% of individuals' mobile phone traces were within the 20 km proximity of their nearest social ties' residential locations. Calabrese et al. (28) defined a co-location event as two mobile phone users being found at the same location at the same time. They found that there was a strong positive correlation between call frequency between two individuals and the frequency of co-location occurrences. Similarly, Bagrow and Lin (88) found that many of locations frequented by a mobile phone user were also frequently visited by their most contacted social ties.

#### *2.2.3.2 Mobility Prediction based on Social Interaction*

Many attempts have been made to exploit the close relationship between the mobility of an individual and that of his social ties to facilitate mobility prediction. Cho et al. (89) constructed a mobility model accounting for travel due to social interaction and reported an order of magnitude better performance than existing mobility models without social network effect. Zhang et al. (90) proposed an algorithm to predict mobile phone users' future movements exploiting their social interplay—a concept capturing social interaction between pairs of users. Their predictor achieved 20% performance improvement over a baseline algorithm without social interplay effect. Domenico et al. (91) showed using the trajectory information of friends of the individual to be predicted greatly improved the prediction accuracy compared to the case where the trajectory of a randomly selected person was used.

## 2.3 MOBILE PHONE DATA

### 2.3.1 Mobile Phone Positioning

One type of mobile phone data that is of particular interest to travel behavior researchers is probably mobile phone positioning data. Mobile phone positioning refers to the attaining of the position of a mobile phone in a cellular network. A cellular network is one that enables mobile phones to communicate with each other; it comprises base stations. The area served by a base station is called a cell (Fig. 2.1). Each cell has a unique cell ID.

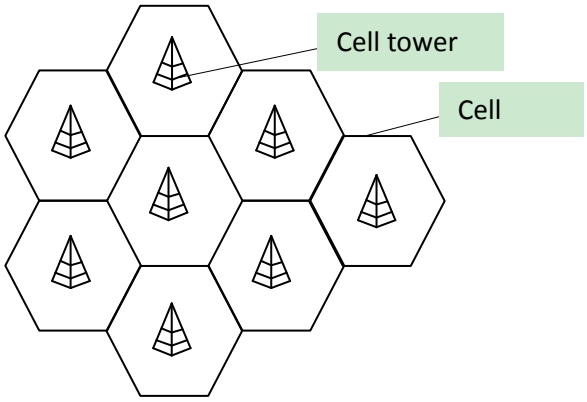


FIGURE 2.1 An example cellular network

There are many approaches of locating mobile phones within a cellular network. Many of them require additional infrastructure to be installed or normal mobile devices to be modified. For instance, in the System for Traffic Information and Positioning (STRIP) project (92), location estimates of mobile phones were obtained by installing monitoring devices along freeway segments to monitor signaling messages exchanged between mobile phones and cellular network. In other examples (93, 94, 95), accurate locations of phones were acquired through built-in GPS receivers in the phones. Yet, these infrastructures and technologies were developed for specific studies and not always available. This dissertation limits its scope to mobile phone

positioning data that is automatically stored by cellular network operators. Hereafter, all mobile phone data in this dissertation refers to this specific type of data.

Cellular network operators don't maintain positions of users at all times due to network performance and bandwidth saving reasons. Positioning is only considered necessary when a user communicates with the network (96). When a user initiates a network connection event (e.g. a voice-call), the cellular network operator needs to know his location in order to determine the cell tower used to channel this event. Therefore, this positioning data only describes users' locations in space when an event occurs. Such data is automatically and passively generated for cellular network operators' own purposes, including collecting billing information and network management.

### 2.3.2 Data Structure

Each time a phone is positioned it generates a single record in a mobile phone data set, i.e. a row in the data set. Each record contains three basic pieces of information: an ID number—a unique number associated with the device generating the record, a location indicating the device's location when this record is generated and a time indicating when the record is generated (Table 2.1). For privacy purpose, the real ID of a device is always encrypted by network operators. The format of location information varies depending on the technique network operators use to perform positioning. The implications of these different technologies on data quality are discussed in the next section. Format of time information can vary as well. Apart from common time formats, Unix time is frequently used in mobile phone data sets.

Mobile data sets can be augmented by other information. Network operators may maintain datasets called Call Detail Records (CDR), in which each record corresponds to a call

activity of mobile phone users. In addition to location and time, each record may also include information about the ID of caller, the ID of callee, duration of the call, etc.

Table 2.1 A Hypothetical Sample Mobile Phone Data Set

| ID | TIME[a] | LOCATION[b] |
|---|---|---|
| 3X35E90 | 1319242582 | 34.044162\|-112.454400 |
| 3X35E90 | 1319242583 | 34.044059\|-112.455550 |
| 3X35E90 | 1319301785 | 34.044392\|-112.453519 |
| 3X35E90 | 1319339560 | 34.040538\|-112.453760 |
| 5YU86I0 | 1315093092 | 33.948195\|-112.170318 |
| 5YU86I0 | 1315093145 | 33.961547\|-112.165304 |
| 5YU86I0 | 1315093169 | 33.977657\|-112.175295 |
| 5YU86I0 | 1315093992 | 34.057944\|-112.178316 |

Note: [a] Time is Unix timestamp—defined as the number of seconds that have elapsed since 00:00:00 Coordinated Universal Time, Thursday, 1 January 1970.
[b] Location is the longitude and latitude coordinates of mobile phones.

### 2.3.3 Spatial Resolution

Depending on the positioning technique adopted by the network operator who provides the data, the spatial resolution of mobile phone data sets can vary.

It is common for network operators to record the location of mobile phones in terms of the cell tower they are currently connected. Yet, in some cases, due to privacy issues, only the ID of the connected tower is provided. Mobile users' traces are, therefore, represented by time-ordered sequences of cell tower IDs, which can be used to infer the topology of cell towers (19). However, since the geographical locations of the towers remain unknown, the spatial resolution

can’t be determined in this case. In other cases, the geographical locations of cell towers are known and can be presented in two alternative ways: the coordinates of the tower or the geographical area where the tower is located. Most of the time, coordinates of towers—their latitude and longitude—are used (56). The spatial resolution of these data sets is determined by the density of cell towers, which varies from as little as a few hundred meters in metropolitan areas to a few kilometers in rural regions. In other words, we could be dealing with an uncertainty level of a few kilometers if the location of users in rural area is considered. In the case where geographical area is used, the study area is first divided into smaller zones (e.g. subprefectures), each of which is served with one or more cell towers. Any phone activity routed through a tower within a zone will result in a record with location represented by the location of this zone (e.g. the centroid of the zone). In this case, spatial resolution of location records highly depends on the size of these zones.

It is also possible for network operators to determine the location of a mobile phone by triangulation, transmission delay from multiple base stations or other more advanced positioning techniques. These techniques can identify the location of phones anywhere in a cell (55) and thus usually result in finer spatial resolution than the cell-tower-based methods, though their accuracy levels of positioning varies as well.

### 2.3.4 Temporal Resolution

Temporal resolution of mobile phone data sets could also vary substantially depending on the specific mobile phone data set. A general categorization of these data sets based on the mechanism triggering records may help to develop some expectations. One type of data sets are Call Detail Records (CDR) as mentioned above. Studies employing this type of data set identify

a ‘burst’ pattern of time intervals between consecutive records/calls: while most subsequent calls are placed soon after a previous call, it is also possible to identify long periods without any call activity. Gonzalez et al. (57) identified an average interevent time as 8.2 hours for 100,000 individuals over a course of six months.

A second type of data sets can be viewed as a superset of the first type of data sets. A record is generated each time an activity is performed on the cell phone, including calling, texting and Internet browsing. It is not surprising that this type of data sets has a finer temporal resolution compared to the first one. Calabrese et al. (48) identified an average interevent time of 260 minutes, which was much lower than that in González et al. (57). They further characterized time interval between consecutive phone activities by its first, second and third quartiles. The authors reported the arithmetic average of the medians as 84 minutes and found the temporal resolution of their data was fine enough to detect changes of location where the user stops for as little as 1.5 hours.

### 2.3.5 Data Processing Techniques

As is the case for almost all raw data sets, mobile phone data sets contain noises, which could have significant implications on study results. In this section, two major types of noises are discussed and some of current data processing techniques used to mitigate the noises are reviewed.

#### *2.3.5.1 Uncertainty in Location Estimation*

As discussed above, advanced positioning techniques, such as triangulation, are capable of estimating the locations of mobile phones within a cell and produce data sets with finer spatial resolution than the cell-tower-based positioning method. In (97), an uncertainty range with an

average of 320 m and median of 220 m was reported. More sophisticated approaches can further reduce localization errors. Zang et al. (98) proposed a technique based on Bayesian inference to locate mobiles in cellular networks. They were able to improve localization accuracy by 20% comparing to a baseline approach with a randomly selected location. For a full review of positioning techniques in cellular network, interested readers are referred to (99, 100). Despite these attempts, uncertainty of location estimation remains. Due to the uncertainty in location estimation, multiple nearby, but distinct, location estimates can occur when a device actually remains at the same location. Thus, these location records need to be aggregated.

There are generally two classes of approaches to aggregate spatial points. One is to impose a grid over the space and aggregate points within each grid cell. In (101), the authors discretized the Seattle area into 1681 1 *km**1 *km* square cells and converted sequences of GPS points to sequences of cells by replacing the coordinates of a point by the index of the cell containing the point. This method highly depends on the layout of the grid (e.g. grid cell size and grid cell shape). As the authors pointed out that, the choice of the layout of the grid was heuristic and they could have chosen a different one. Ye et al. (102) described another problem of this grid-based technique: grid boundaries could be problematic when points corresponding to the same place falls in different grids.

The other class of approaches to aggregating spatial points is through clustering. Clustering-based approaches allow points to be aggregated with arbitrary shape and oftentimes require a distance threshold as input. Ye et al. (102) aggregated a sequence of points into one location if 1) the temporal difference between the first point and the last point was more than 30 minutes and 2) all the points were within a range of 200 meters. Similarly, in a series of studies

with mobile phone data from the Boston area (34, 48), sequences of points were fused into one location if the distance between any two of them was less than 1 km.

General procedure of clustering-based approaches can be summarized as following. First, it starts with the series of location records for an individual ordered by time stamps, denoted as $= \{l_{t_1}, \dots, l_{t_n}\}$. Second, the first location record $l_{t_1}$ is chosen to be the center of the first cluster and the distance between the second location record $l_{t_2}$ and $l_{t_1}$ is calculated. If the distance is less than a threshold $k$, then $l_{t_2}$ is fused into this cluster and the cluster center is updated as the geometric center of $l_{t_1}$ and $l_{t_2}$. If the distance is greater than $k$, $l_{t_2}$ becomes the center of a new cluster. Third, the second step is repeated for all the remaining location records $\{l_{t_3}, \dots, l_{t_n}\}$ until all the points are assigned to a cluster. All the points within a cluster are then analyzed as a virtual location for subsequent analysis. Fig. 2.2 illustrates this procedure graphically.

The distance threshold in above studies was determined, to a large extent, heuristically. It is recommended that, if clustering-based approaches are to be adopted, sensitivity analysis needs to be performed in order to fully evaluate the implications of different distance thresholds on location detection.

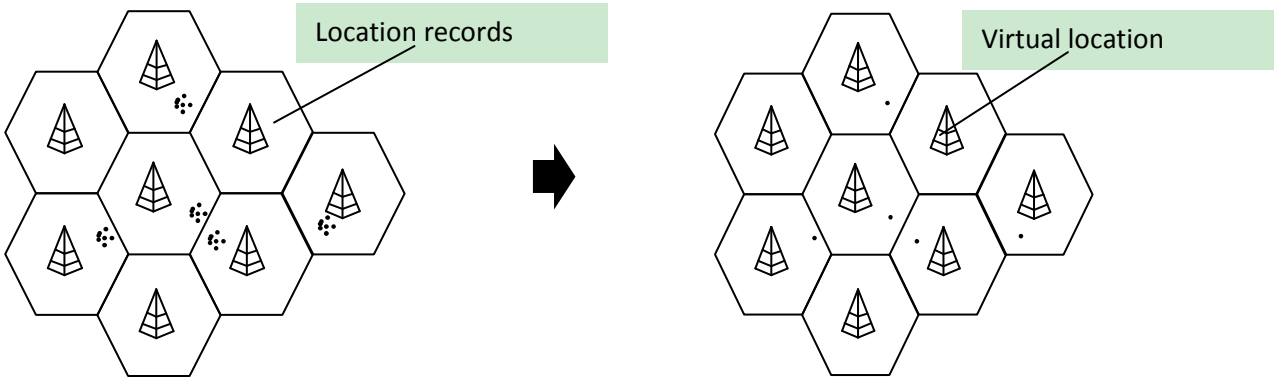


FIGURE 2.2 Clustering location records

*2.3.5.2 Oscillation*

At any given location in a cellular network, there may be several cell towers whose radio signals reach a device. If these multiple cell towers have similar signal strengths, the connection of a device may hop between multiple towers even when the device is stationary. In this case, it may appear that the user travels for several kilometers in just a few seconds. This phenomenon is known as oscillation in a cellular network. Fig. 2.3 illustrates the potential impacts of this oscillation phenomenon on the detection of a device's location. A device is on the boundary of cell A and cell B and the signal strengths received by this device from tower A and tower B are equal. This device can be registered to either tower A or tower B depending on the real-time traffic through these two towers. When it is registered to tower A, its location will be recorded as location A. Similarly, its location will be recorded as location B when it is handed over to tower B. Distinct location records—location A and location B—resulting from oscillation need to be consolidated. A few methods have been proposed to address this oscillation problem.

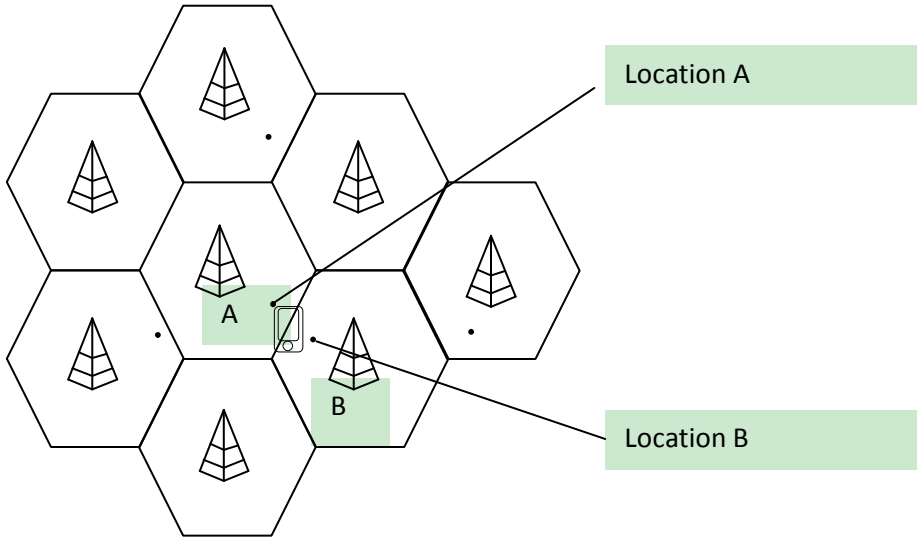


FIGURE 2.3 Oscillation in a cellular network

Iovan et al. (103) proposed a speed-based method: if location B is recorded in the middle of two location A records and the switch speed from location A to location B is larger than a predetermined threshold, oscillation was detected. This method is based on the observation that oscillation results in a location change characterized with an abnormally high speed. Yet, a critical question involved in this method is the choice of a speed threshold that distinguishes 'normal speed' and 'abnormal speed'. On the other hand, a pattern-based method has been applied in some studies. This method recognizes the unique pattern in location updates associated with oscillation—frequent switches between pairs of locations. Lee and Hou (104) identified the occurrence of oscillation as each time three consecutive mutual switches between a pair of locations were observed. Once oscillation is identified, all the locations involved in these switches are replaced with that location in the pair with which the user has been associated most of the time. A similar method was also adopted by (19).

Procedure used to perform this pattern-based method in (104) can be described as following. A sequential scan starts from the beginning of location records of a mobile phone user ordered by time stamps. If a sub-sequence of location records contains mutual switches between two locations for at least three times, such as $\{X_{t_1}, A_{t_2}, B_{t_3}, A_{t_4}, B_{t_5}, Y_{t_6}\}(t_1<t_2<\cdots<t_6)$, oscillation is considered present. This sub-sequence is then updated so that all location records in this sub-sequence would indicate just one location—the one which the user has been associated with most of the time. In the same sub-sequence above, if the user is found to be associated with tower A for more time than tower B, then location B is replaced by location A, which results in an updated sub-sequence as $\{X_{t_1}, A_{t_2}, A_{t_3}, A_{t_4}, A_{t_5}, Y_{t_6}\}$.

This pattern-based method has the risk of mistaking the actual movements of a user who frequently travels between two locations for oscillation. Here, it is considered that a combination

of the speed-based and pattern-based approaches may render more reliable results: firstly, sub-sequences seemingly resulting from oscillation are detected based on pattern-based approaches; secondly, switching speeds between pairs of locations is determined for each sub-sequence; lastly, sub-sequences are only updated if the switching speed is beyond a speed threshold as determined in speed-based approaches.

## 2.4 PROSPECTS AND ISSUES OF MOBILE PHONE DATA

### 2.4.1 Relative Advantages

As an alternative approach for travel data collection, mobile phone data has several major advantages over travel diary/GPS tracking data.

1) Much larger sample size: Studies show that traditional travel surveys usually have a sample size less than 10,000 (105), while, with mobile phone data, a sample size of ~1 million is not uncommon.

2) Longer duration: Theoretically, mobile phone data can be collected for as long as necessary. Currently, mobile phone data sets spanning from a few months to several years have been found in empirical studies (28, 56). In contrast, most of the large-scale (regional or national) travel diary data sets are of one-day only with a few being two days; those studies that last more than a few days are typically smaller-scale unrepresentative ones (106). Recent advancements in GPS tracking technology have made feasible the collection of multi-day travel data (107, 108). Yet, rarely can one identify GPS data sets lasting over one month.

3) Cost-effectiveness: Travel surveys on average cost $487,000 (105) with multi-day surveys costing more. On the contrary, mobile phone data is automatically collected by the

cellular network operators with no additional cost, though there will be cost of purchasing the datasets from the network operators.

4) No human errors: Human errors can be introduced into travel diaries in many ways (109). During mobile phone data collection, subjects have no active participation, which precludes the possibility of any human error.

5) No non-response: Non-response rate is one of the major concerns in travel surveys. This problem becomes more salient in multi-day travel data collection as subjects are required to record and enter information over an extended period (109). Subjects in mobile phone data sets, however, are automatically included as cellular network subscribers. Thus, there is no concern of non-response.

6) No fatigue/attrition: One of the most important concerns with multi-day travel data sets is reporting fatigue, evidenced as decreased number of days with at least one trip and/or decreased number of trips reported over time (110). In longitudinal studies, attrition can be a severe problem, with participants' opting out in the middle of a survey. While reporting fatigue always occurs in multi-day travel diary, GPS tracking data can suffer from similar problems when participants forgot to turn on or charge GPS devices later in the study period. Without being actively involved, subjects in mobile phone data sets don't experience fatigue.

### 2.4.2 Unresolved Issues

Though the advantages of mobile phone data sets are evident, several issues remain in applying mobile positioning data to study travel behavior.

1) Proximity of mobile phones: Studying travel behavior with mobile phone data implicitly assumes that individuals would always carry their devices around and thus the positions of these devices serve as a reasonable proxy of the users' locations. Studies have

recently raised questions about this assumption. In (111), the authors categorized the proximity of mobile phone to its user into three levels: within arm's reach (1-2 meters), within the same room (5-6 meters), and unavailable (beyond 6 meters). Results show that proximity levels varied substantially for different users and under different circumstances (e.g. in or out of home). For one of the study participants, his mobile phone was unavailable for more than 70% of the time. Yet, the authors also showed individuals tended to keep their phones close when travelling (out of home). Clearly, if mobile positioning data is to be used for large-scale, regional travel behavior studies, more studies understanding how different population segments carry and use their mobile phones are needed.

2) Multiple mobile phones: Surveys (22) have shown that in many countries, the penetration rate of mobile phones is over 100%, which means some individuals are likely to carry multiple devices. Implicitly in many existing studies is the assumption that each device uniquely represents one individual. These individuals are over-represented in a mobile positioning dataset and may overshadow the behaviors of others. Studies are needed to understand this particular segment of the population in terms of their size as well as their carrying and phone use behaviors.

3) Penetration rate: Mobile phone data sets can suffer from unrepresentativeness depending on mobile phone penetration rate in the study population. Though this may not seem to be a problem in developed countries, mobile phones are far from ubiquitous in many developing countries. Recent news shows that the mobile phone penetration rate just reached 55% in Rwanda (112). Individuals who don't own mobile phones are precluded in studies. It is expected, though, this will be resolved as the penetration rate keeps rising throughout the world. Second, depending on the cellular network operator(s) who provides the positioning data, non-

subscribers are precluded and thus underrepresented. As the biggest cellular network operator in U.S., Verizon only holds a market share of 32% (113) and there are dozens of other operators in the U.S. Little is known about whether there exists some systematic difference in the travel behaviors of subscribers with different cellular network operators.

4) Sample selection: It is common for researchers to select a study sample from all the subscribers included in a raw mobile phone data set provided by the network operator, and this selection can be non-random and renders the final sample unrepresentative. As an example, in (56), a sample of mobile phone users who made at least one call every two hours was selected. Recent studies (103, 114, 115) show that user mobility had a strong correlation with phone usage: more active users are more mobile. Therefore, sample selection based on phone usage would potentially result in an overestimation of mobility levels. On the other hand, some studies (103) also suggested that some mobility measures seem to be immune to this sampling bias. In summary, great caution should be exercised when mobility information derived from mobile phone data are to be generalized to the general population.

5) Positioning accuracy: Mobile phone data has a much lower positioning accuracy level compared to GPS tracking data (usually with an error range of less than 10 meters). Mobility measures derived from mobile phone data have been compared to those from GPS tracking data (116) and the validity of these measures differs: average daily travel distance estimated with mobile phone data is less than that derived from GPS tracking data, while frequent activity locations detected with both data sets are highly consistent. More studies aimed at cross-checking mobility measures from mobile phone data are much needed (117).

6) Socio-demographic information: Mobile phone data usually doesn’t contain users’ socio-demographic information. If the research interest is to explain mobility measures derived

from mobile phone data with socio-demographic variables, mobility measures can be aggregated to a geographic level where socio-demographic information is publicly available. In (97), individuals' daily trip lengths derived from mobile phone data were aggregated to block-groups level and associated with socio-demographic information from U.S. Census. Many more studies are needed to check the validity of such procedures and comparability across regions. It is worth to note that, even though such procedures can be validated, individual level socio-demographic information remains unavailable as required in conventional disaggregate travel behavior modeling.

7) Privacy: Usually privacy protection is achieved by researchers receiving an anonymous data set from cellular network operators. Also, research results are supposed to be published at aggregated level (54). Researchers also have the choice to adopt an 'opt in' policy so that individuals' permission is guaranteed before their data is used for research purpose (55). This 'opt in' policy would potentially reduce sample size. In (32), the authors asked 576 individuals' agreement for monitoring their phones for research purpose. 231 of them agreed and the main reason for refusal was not privacy related (but because they don't have a contract with a specific cellular network operator). Only 10% showed a serious concern of surveillance.

## 2.5 CONCLUSIONS

Data generated from the ubiquitous mobile phone system provides an alternative data source for research and is yet to be explored by the travel behavior research community. Advantages of mobile phone data over travel data collected through traditional approaches are evident in many aspects. Certainly, mobile phone data is not without limitations. Among others, its coarse granularity in space and time prevents us to obtain the ground-truth of users' trajectories. There

is vast uncertainty of a user's whereabouts when he is not communicating with the network (96). Therefore, mobility information extracted from mobile phone data still needs to be carefully evaluated depending on specific applications. Despite these unresolved issues, mobile phone data has seen its success in many travel behavior studies and it possesses enormous potential to be continuously exploited by the travel behavior research community. A few possible directions are discussed below.

First, validation of techniques used to infer behavioral measures from mobile phone data is in imperative need. Current techniques used to derive behavioral measures from mobile phone data are largely exploratory and requires validation. Existing validation is mostly implemented by comparing the derived behavioral measures to statistics recorded by independent sources. For instance, Becker et al. (118) proposed a new technique to determine the commute route taken by individuals into Morristown, New Jersey based on their mobile phone traces. A correlation coefficient of 0.77 between their traveler count on each commute route and the traffic counts published by the New Jersey Department of Transportation was reported. Though this number seems promising, discrepancy remains. And rarely the cases that further insights and discussions are provided for the discrepancy: 1) what are the underlying causes of this discrepancy? 2) what implications this discrepancy have on final results? Studies looking at these questions should provide us with more information on techniques we use to derive behavioral measures from mobile phone data and the generalizability of our results. Further question arises in the case that no reliable reference is available for validation. In (73), for instance, the authors noted that there was a temporal gap between individuals' residential location derived from mobile phone data and those as a reference and no other better reference can be located. This necessitates an

accumulation of validation studies so that the validity of certain techniques can be assured even without references.

Second, more research should be directed to extract higher-level behavioral knowledge from mobile phone data. As mentioned above, many mobile phone data sets only contain basic information as location and time. Yet, location and time constitute only a small part of a person's state (119). As travel behavior researchers, we are interested in other behaviors, such as travel mode, and the context of such behaviors, such as the activities performed. These kinds of information are not explicit in mobile phone data sets. Some early efforts have been made to infer activity information from mobile phone data. In (120) and (121), the authors identified the most probable activity type associated with a specific location based on surrounding points of interest. However, the activity type derived was rather general (i.e. eating, shopping, entertainment and recreational) and no other activity information, such as activity duration, was available. Discovery of higher level knowledge from location history requires dedicated techniques worth more detailed studies (102).

A related issue is the combination of mobile phone data with other travel survey data. It has been recognized that mobile phone data will most likely be complemented by other travel data in performing travel behavior analysis. Yet, it is also desirable to exploit information in mobile phone data to minimize complementary data collection efforts. As pointed out in previous section, one major drawback associated with many mobile phone data sets is the lack of socio-demographic information. Previous studies addressed this problem by integrating socio-demographic information from census (97). In fact, recent studies have shown that it is possible to infer some socio-demographic variables based on information contained in mobile phone data (122, 123). These efforts will definitely expand the scope of application of mobile phone data.

Third, mobile phone data provides us with the opportunity to revisit many research topics that have been hampered by the lack of data. Here, a couple of examples is provided here in the hope of stimulating further discussions. For starters, mobile phone data can facilitate the study of social influence on travel behavior. There has been increasing interest in integrating social dimensions in understanding travel behavior in recent years (124). It is hypothesized that social interaction is an underlying cause of travel behavior. However, the lack of data on individuals' social network structure has inhibited researchers from uncovering a reliable link between social network and travel behavior (125). Since phone communication serves as a primary vehicle of information diffusion in social network, the communication information contained in mobile phone data sets can be used to infer social network structure, and is expected to significantly enrich our knowledge on the social context of travel behavior.

Another research area that can benefit from the use of mobile phone data is the study of dynamics of travel behavior. Previous studies on the dynamics of travel behavior are generally performed with two types of data sets: multi-day data sets and panel data sets (9). Multi-day data sets contain records of behavior for a period ranging from a couple of days to a few weeks. Panel data sets contain behavioral data of a sample that is repeatedly surveyed in multiple waves at distinct points in time, usually months or years apart. Both have limited ability in capturing the complexity of temporal dynamics in travel behavior. Multi-day data sets are limited in their length, which prevents us to discover behavior patterns relying on long-term observations (126). Panel data sets suffer from its discontinuity, which could leave out many critical moments in behavior evolution (127). Mobile phone data, on the other hand, can be continuously collected for a prolonged time period and allow us to identify infrequent travel behaviors (e.g. long-

distance travels) and understand the behavioral path towards behavioral changes. In (35), mobile phone data was used to identify individuals' residential locations and approximately 5% of the population in Estonia was found to change their residential locations seasonally. Very recently, Järv et al. (128) has applied mobile phone data to study the monthly variability in individuals' activity space. Aided by mobile phone data, Beckor et al. (129) was able to characterize long-distance travel of the population of Israel at national level.

In summary, mobile phone data opens new opportunities for travel behavior research and has already stimulated enormous research interest in mobility behavior from many other disciplines, as evidenced by the number of studies reviewed in this chapter. It is our hope that this dissertation can help in synchronizing travel behavior research with research conducted in the age of 'big data' and foster dialogues between travel behavior research and other domains.

# CHAPTER 3

# DATA

## 3.1 DATA OVERVIEW

The mobile phone data set used in this dissertation contains 128,412,557 sightings generated by a sample of 120,435 individuals during the time period between September 1st and October 31st, 2011. Each time a phone communicates with a tower, including calling, texting and Internet browsing activities, a sighting is generated and it becomes a row in our data set. Each sighting is composed of an ID number—a unique number associated with the device generating the sighting, a location estimate indicating the device's location when this sighting is generated and a time indicating when the sighting is generated. For privacy purpose, the real ID of devices has been encrypted and the ID available in our data set is a random combination of numbers unique to each device. The estimated location is in the format of longitude and latitude coordinates. The time stamp of each sighting is in Unix time.

### 3.1.1 Number of Sightings

Fig. 3.1 shows the variations in daily total number of sightings generated by the sample. Though the existence of variations over time is evident, daily total number of sightings fluctuates around 2,000,000. Daily total number of sightings exhibits a prominent weekly cycle. The total number of sighting stays relatively stable (i.e. no linear trend) from Tuesdays to Fridays and peaks on Saturdays. After a significant decrease over Sundays, the total number of sightings hits its minimum on Mondays corresponding to the dates of September 12, September 19, September 26, October 3, October 10, October 17 and October 24, 2011.

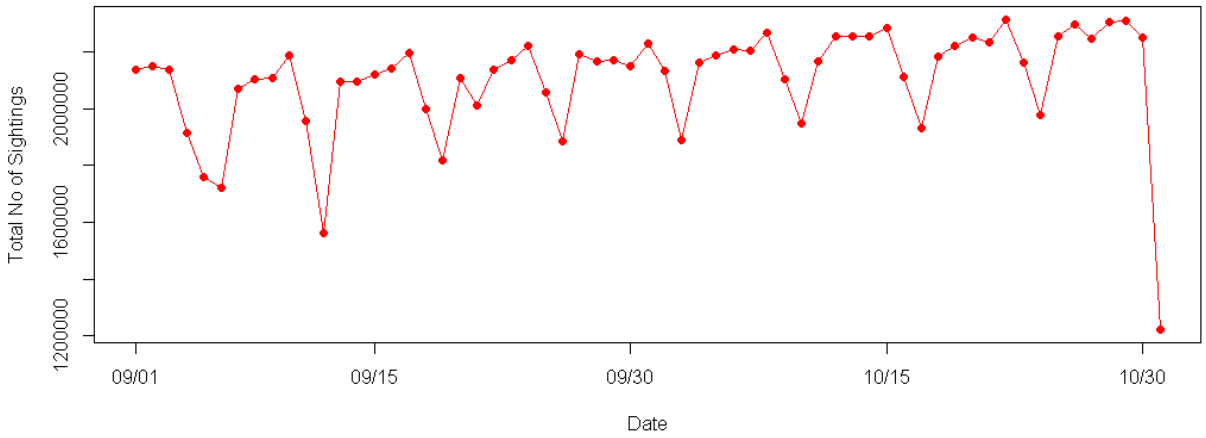


FIGURE 3.1 Variations in daily total number of sightings

Fig. 3.1 reveals little about the distribution of daily total number of sightings among individuals. Therefore, Fig. 3.2 is presented.

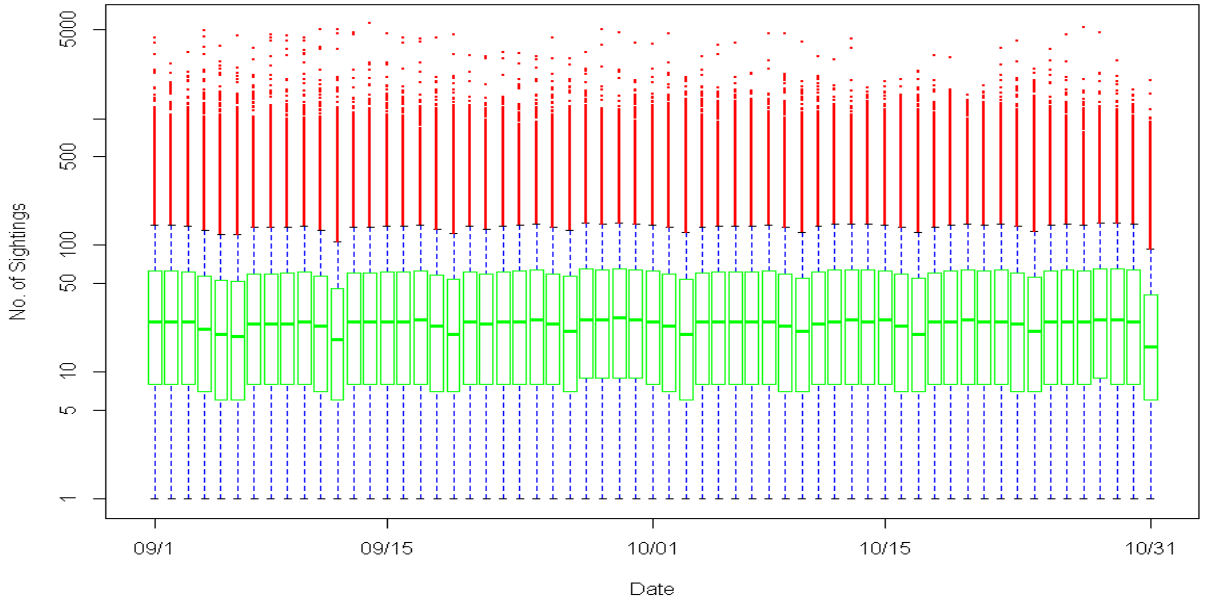


FIGURE 3.2 Boxplot of daily number of sightings

Fig. 3.2 is the box-and-whisker plot showing the distribution of daily total number of sightings among individuals. Green boxes cover the range between the first quartile and the third quartile with the median value clearly marked in the middle. The blue whiskers extend to 1.5 times the range between the first quartile and the third quartile and the red dots are outliers. In general, it shows that daily total number of sightings is unevenly distributed among individuals and contains significant amount of heterogeneity: most of individuals generated a few dozens of sightings each day, while some generated as few as less than ten sightings. In addition, there are a small number of individuals who generated extreme high number (thousands) of sightings each day, which is indicated by those outliers.

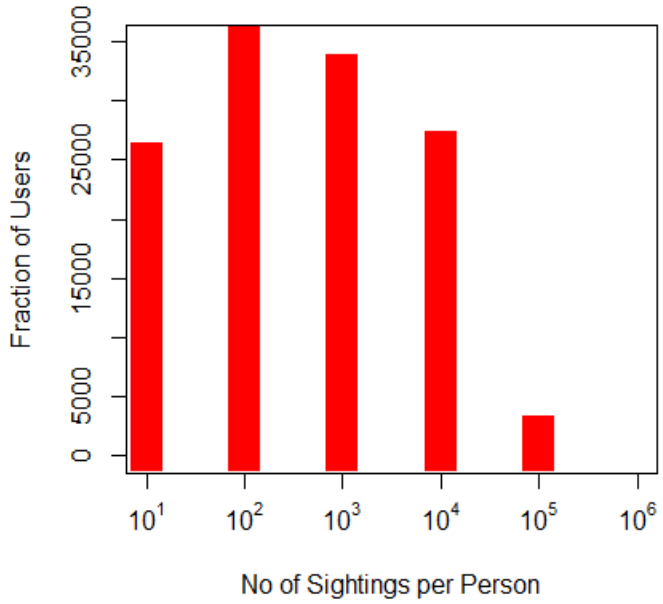


FIGURE 3.3 Distribution of total number of sightings

Fig. 3.3 shows the distribution of total number of sightings in two months. Similar to what is observed for daily number of sightings, total number of sightings also varies significantly

among individuals: 21% (25,054) of the sample generated fewer than 10 sightings; 29% of the sample (34,939) generated a total number of sightings ranging from 10 to 100 and 49% of them (58,463) generated sightings somewhere between 100 and 10,000; only a small portion (2%) generated more than 10,000 sightings.

### 3.1.2 Total Number of Sightings Decomposition

The difference in the total number of sightings can result from two sources: difference in average daily number of sightings and/or difference in the number of days observed. Therefore, for each individual, two quantities, namely the average daily number of sightings and the number of days observed, are subsequently investigated.

Fig. 3.4 shows the distribution of the number of days observed.

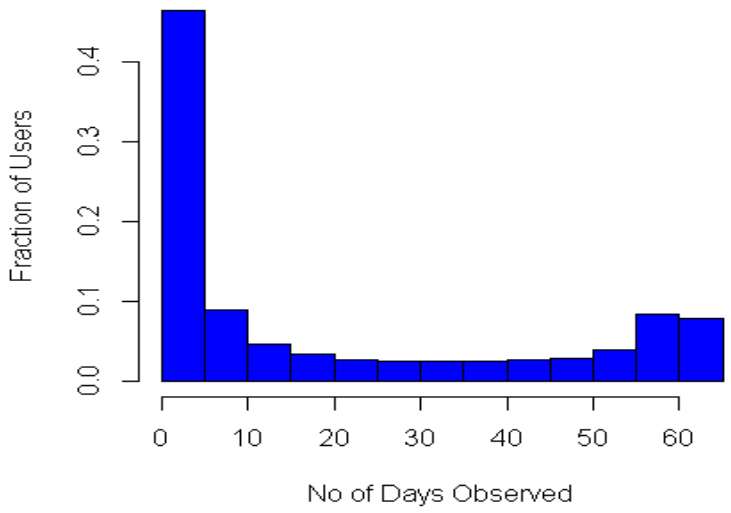


FIGURE 3.4 Distribution of number of days observed

About 47% of the sample (56,113 individuals) were only observed for fewer than 5 days. This number is followed by those observed for 5 to 10 days (10,847 or 9% of the sample) and more than 55 days (19,619 or 16% of the sample). The rest of the sample is almost evenly distributed over the span from 10 to 55 days. In summary, this figure shows that individuals differ substantially in terms of the number of days observed.

If difference in the total number of sightings is completely attributable to the difference in the number of days observed (i.e. individuals generate the same average daily number of sightings), a perfect correlation (with a correlation coefficient as 1) between the total number of sightings and the number of days observed should be observed. However, the actual coefficient computed is 0.54, which means that the difference in the total number of sightings is more likely to be a convolution between the difference in the number of day observed and the difference in average daily number of sightings. Consequently, distribution of average daily number of sightings is examined in Fig. 3.5.

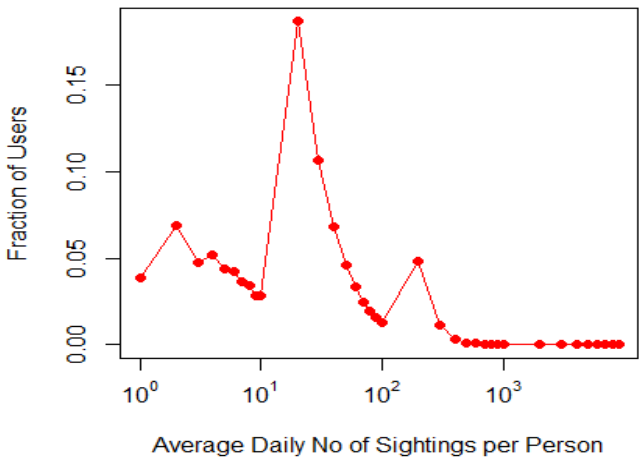


FIGURE 3.5 Distribution of average daily number of sightings

On average, 62,919 individuals (52% of the sample) generated a number between 10 and 100 of sightings every day. There are 49,541 individuals (41% of the sample) who generated fewer than 10 sightings per day. Only 7,975 (6% of the sample) individuals generated a daily number of sightings more than 100, among whom 16 individuals generated more than 1,000.

### 3.1.3 Time Intervals

Statistics of the number of sightings reveal little about the time interval between consecutive sightings. In order to reconstruct individuals' trajectories with sightings, sightings should have a sufficiently fine temporal resolution. Therefore, time intervals between consecutive sightings are examined.

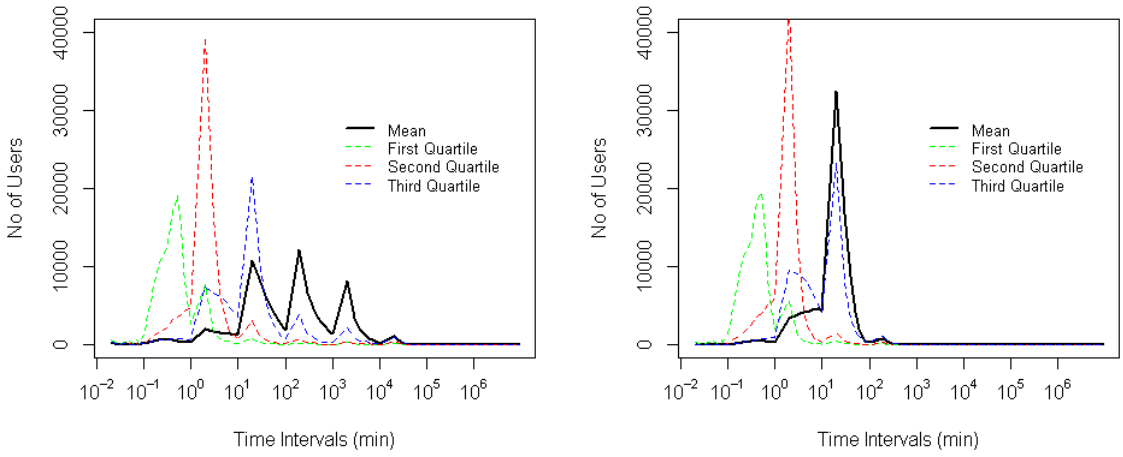


FIGURE 3.6 Time intervals between consecutive sightings (left panel: all sightings; right panel: daytime (6 a.m.-12 a.m. the next day) sightings)

For each individual, the mean and first, second, third quartiles of time intervals are computed and the distributions of these statistics are graphically presented in the left panel in Fig. 3.6. The figure shows that the first, the second and the third quartiles peak at less than 1 minute,

between 1 and 10 min, and 1 hour, respectively, while the means seem to spread over the range between 10 min and a day (1,440 min). The distinction between the distribution of the means and that of the medians (i.e. the second quartile) suggests that the distribution of time intervals are positively skewed, which is characterized by some very long time intervals. Long time intervals would greatly hamper the reconstruction of individuals' trajectories. It's suspected that most of these long time intervals occur during night time, when devices are relatively inactive. In order to confirm this speculation, sightings from 12 a.m. to 6 a.m. of each day are removed and the distribution of the same of set of statistics (mean, first, second and third quartiles) are recalculated and shown in the right panel in Fig. 3.6. The distribution of means greatly shifts to the left and is characterized with a single peak at around 1 hour. Though the distributions of other statistics don't alter much, their overall values become smaller with those bumps in the tail of the original distribution removed.

## 3.2 DATA PRE-PROCESSING

Of interest in this dissertation is the variability of activity locations. In the following, a procedure is presented to extract individuals' activity locations from location estimates in our mobile phone data set.

### 3.2.1 Clustering of Sightings

Due to uncertainty in the location estimation, multiple distinct location estimates can occur when a device actually remains at the same location. Therefore, nearby location estimates are clustered following the practice of previous studies (48, 130). Firstly, all the distinct location estimates for an individual, denoted as $L = \{l_1, \dots, l_N\}$ are extracted ($N$ is the total number of

distinct location estimates). Secondly, the first location estimate $l_1$ is chosen to be the initial center of the first cluster and the distance between the second location estimate $l_2$ and $l_1$ is calculated. If the distance is less than 1 $km$, then $l_2$ is fused into this cluster and the cluster center is updated as the geometric center of $l_1$ and $l_2$. If the distance is greater than 1 $km$, then $l_2$ becomes the center of a new cluster. Thirdly, a similar procedure to that in the second step is repeated for all the remaining location estimates $\{l_3, \dots, l_N\}$. That is: the distances from location estimate $l_i$ from each of the existing cluster centers are calculated successively. If a distance is less than 1 $km$, then $l_i$ is fused into this cluster and the cluster center is updated as the geometric center of all the points included. If all distances between $l_i$ and cluster centers are greater than 1 $km$, than $l_i$ becomes the center of a new cluster. Lastly, each location estimate is replaced with the cluster center of its associated cluster. Each cluster center is considered to be a virtual location where those sightings whose estimated locations fall into this cluster are generated and these virtual locations are used in the following analysis.

The 1 $km$ threshold is chosen to take account for the location estimation errors in the data set. Ideally, clusters should be apart by a distance that is larger than twice the location estimation error so that location estimates in a distinct cluster are less likely to be estimates of a single location. In order to observe the change in the distance between clusters with respect to increasing threshold distance, the minimum distance between clusters for each individual is calculated for 500 m, 1000 m and 1500 m as threshold distance for clustering. Results are presented in Table 3.1, along with the total number of clusters.

The minimum distance between clusters increases as the threshold distance becomes larger. For the 500 m threshold, a quarter of the sample has a minimum distance between clusters less than 280 m, which is comparable to the location estimation error in my data set with an

average of 320 meters and a median of 220 meters (130). Larger threshold distance produces better-separated clusters. For the 1000 m and 1500 m thresholds, a large portion of the sample (more than 75%) have a minimum distance larger than twice the location estimation error. Distances between clusters resulted from both 1000 m and 1500 m thresholds are satisfactory. On the other hand, the number of clusters decreases as the threshold distance increases: distinct clusters are merged together with larger threshold distance. In order to better preserve the spatial pattern in the data, a smaller threshold distance—1 km threshold is used for following analysis.

Table 3.1 Cluster Distance and Number of Clusters

| | | Min. | 1st Quartile | Median | Mean | 3rd Quartile | Max. |
|---|---|---|---|---|---|---|---|
| 500m | No. of clusters | 1.00 | 5.00 | 20.00 | 47.32 | 65.00 | 926.00 |
| | Min. distance between clusters (m) | 2.10 | 280.00 | 451.50 | 836.20 | 585.70 | 71240.00 |
| 1000m | No. of clusters | 1.00 | 4.00 | 13.00 | 26.22 | 37.00 | 382.00 |
| | Min. distance between clusters (m) | 2.94 | 562.00 | 919.10 | 1577.00 | 1196.00 | 73150.00 |
| 1500m | No. of clusters | 1.00 | 3.00 | 10.00 | 18.20 | 26.00 | 231.00 |
| | Min. distance between clusters (m) | 4.19 | 870.60 | 1438.00 | 2268.00 | 1864.00 | 83030.00 |

### 3.2.2 Oscillation

At any given location, there may be several cell towers whose radio signals reach a device. If cell towers have approximately equal signal strengths, a device may hop between cell towers even when it is not moving. In this case, it may appear that the user travels for several kilometers in just a few seconds. This phenomenon is known as oscillation in cellular network. For instance, if a device is assumed to be on the boundary of cell A and cell B, the signal strength received by this device from tower A and tower B is approximately equal. This device can be registered to either tower A or tower B depending on the real-time traffic through these two towers. When it is

registered to tower A, its location will be recorded at location A. Similarly, its location will be recorded as location B when it's handed over to tower B. Distinct location records—location A and location B—resulting from oscillation need to be consolidated. Otherwise, it would lead to an overestimation of the total number of activity locations visited. This problem is addressed in this dissertation by a two-phase approach: detecting oscillation location series and updating oscillation location series.

*3.2.2.1 Detecting oscillation series*

When oscillation occurs, a unique pattern in location records is observed. Specifically, there is a series of location records which consist of frequent switches between a pair of locations. Take location A and location B as an example. A series would appear like $\{location\ A_{t_1}, location\ A_{t_2}, location\ B_{t_3}, location\ A_{t_4}\ , location\ B_{t_5}, location\ A_{t_6}, location\ A_{t_7}\}(t_1<t_2<\cdots<t_7)$. This type of series is termed as an oscillation series in this dissertation. Oscillation series are prevalent in this dataset. The number of trajectories (out of 120,435) contains 2-time switches, 3-time switches and 4-time switches between any pair of locations are found to be 83,463, 65,243 and 57,679, respectively.

A heuristic rule, similar to that in Bayir et al. (19), is used to detect oscillation series: if a series of location records is observed to be switching between two locations for at least 3 times, it is qualified to be an oscillation series. In the same example, we observe the switch always happens between location A and location B and there are totally 4 (>3) switches at $t_2 - t_3$, $t_3 - t_4, t_4 - t_5, t_5 - t_6$. Therefore, this series of location records is an oscillation series. In addition to the 3-time switches threshold, 2-time switches and 4-time switches thresholds in identifying oscillation series are also tested for sensitivity analysis purpose. Daily number of

activity locations visited resulting from different thresholds is discussed in the next section (Section 3.2.3).

*3.2.2.2 Updating oscillation series*

After an oscillation series is identified, it's updated so that all the location records in this series would indicate just one location—that one which appears more frequently in the series. In the same oscillation location series above, location A is observed for 5 times and location B is observed for 2 times. Thus, in this series, location B is replaced by location A, which results in a series as $\{location\ A_{t_1}, location\ A_{t_2}, location\ A_{t_3}, location\ A_{t_4}\ , location\ A_{t_5}, location\ A_{t_6}, location\ A_{t_7}\}$.

### 3.2.3 Activity Location Selection

Only activity locations are of interest in this dissertation. It's proposed that stay duration at a location must exceed a threshold in order to qualify a location as an activity location. Stay duration at a location is calculated as the time difference between the first and the last record in a sequence of consecutive records generated at this location. Following Bayir et al. (19), the stay duration threshold distinguishing activity location from non-activity location is set to be 10 min. In Kim and Kwan (131), the authors also argued that 10-min was the minimum duration required for the meaningful participation in any activity.

In order to get a taste of the general pattern of individuals' activity location choices, the distribution of the number of locations visited for various time periods is presented in Fig. 3.9 (left) with a comparison from Song et al. (56) (right).

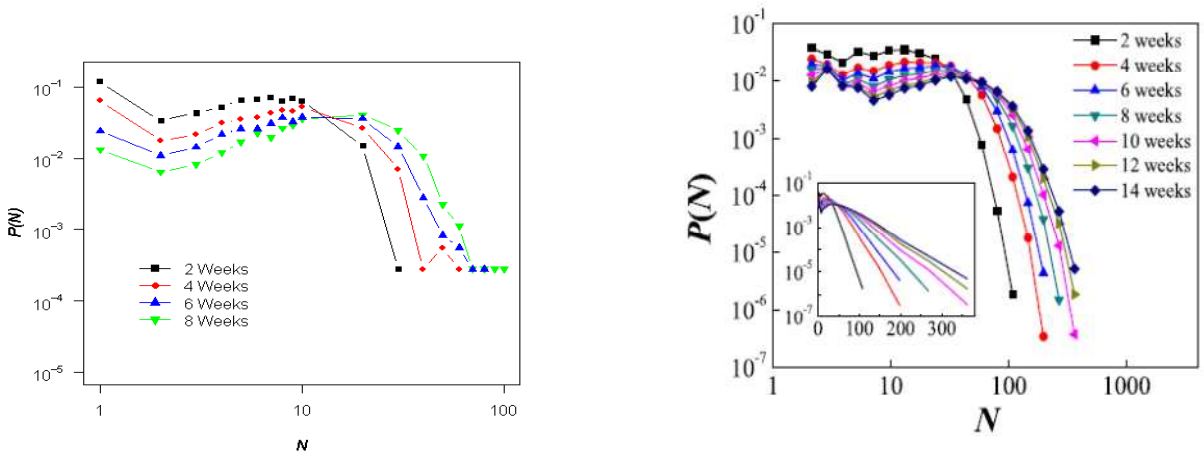


FIGURE 3.7 Distribution of number of locations *N* visited for various time periods (left panel: data used for this dissertation; right panel: from Song et al. (2010))

The general trend in the two figures is consistent: most of the people visited only dozens of locations during the 8 weeks observed and the probability converges and becomes saturated over time. Yet, curves based on our data (left) shift to the left, i.e. the number of location visited is smaller. Two steps in our data processing procedure explain this difference. First, our locations are actually activity locations with stay duration of at least 10 minutes and, intuitively, the number of activity locations is fewer than the number of all the locations (including both activity locations and transient locations) recorded in Song et al. (56). Second, the oscillation problem does not appear to be explicitly addressed in their paper, which could lead an overestimation of the total number of locations. It is also worthy to note that, in Song et al. (56), a location refers to the cell tower a device connected to, while, in this dissertation, after the preprocessing (clustering and addressing oscillation), it refers to an activity location.

Besides the total number of activity locations, also calculated is average daily number of activity locations visited by each individual. Statistics are presented for different thresholds in

identifying oscillation series (i.e. 2-time, 3-time and 4-time switches), as well as for the default case where oscillation effect is left unaddressed.

Table 3.2 Daily Number of Activity Locations

| | Min. | 1st Quartile | Median | Mean | 3rd Quartile | Max. |
|---|---|---|---|---|---|---|
| Default | 1.00 | 1.00 | 1.50 | 2.10 | 2.50 | 20.00 |
| 2-time | 1.00 | 1.00 | 1.33 | 1.78 | 2.03 | 21.17 |
| 3-time | 1.00 | 1.00 | 1.40 | 1.85 | 2.16 | 20.14 |
| 4-time | 1.00 | 1.00 | 1.41 | 1.86 | 2.20 | 19.22 |

In general, numbers presented in Table 3.2 are lower considering the average 3.79 trips per person reported in 2009 National Household Travel Survey (132). There are a few possible reasons. First, the scope of dissertation is limited to daytime travel (6 a.m. to 12 a.m. the next day). So nighttime trips are excluded. Second, for those individuals having sparse sightings, activity locations visited between consecutive sightings may not be captured, which resulted in lower number of location visited. Third, stay duration at one location is calculated as the time difference between the first sighting at this location and the last one. Since the first sighting and the last sighing at a location usually occur during the stay, the stay duration calculated is generally shorter than the actual stay time. It is possible some activity locations can be ruled out if actual stay time is underestimated by the calculated stay duration.

As expected, daily number of activity location visited decreases after oscillation effect is removed. Moreover, it increases along with the number of switches. It is suspected that, with fewer number of switches, more actual trips are mistaken for oscillation effects. Consider a simple trip chain with 2 activity locations: work place—lunch place—work place. Computed number of activity location visited would be 1 with 2-switches threshold as opposed to 2 with 3-

switches or 4-switches thresholds. So a higher number of switches is considered to be more effective in detecting oscillation effect. Yet, from 3-switches to 4-switches thresholds, change in daily number of activity locations visited is negligible. Therefore, the 3-time switches threshold will continue to be used in this dissertation.

# CHAPTER 4

# TIME-OF-DAY DEPENDENCE OF LOCATION VARIABILITY

## 4.1 INTRODUCTION

Although the existence of variability in individuals' travel is well recognized (2, 11), Our current understanding of variability in travel behavior remains limited. Variability in urban travel behavior has received little attention in the literature, primarily because most data sets used for analyzing and modeling urban travel comprise information for just a single day for each sampled individual and thus preclude the examination of variability (12). Despite the spatial nature of human travel, the limited research on variability in travel behavior has been largely focusing on non-spatial aspects, such as the number of daily trips, daily travel distance and daily travel time (106). Lack of knowledge on variability in spatial dimensions (e.g. activity locations) leads to a potential gap in the current thinking on the relationship between activity-travel behavior and the consumption of urban space (106). For instance, an individual can be observed to make the same number of trips on two days, though to completely different sets of activity locations. This dissertation makes an effort to quantify the variability in individuals' spatial behaviors by examining the spatial variability of activity locations, i.e. the extent to which individuals either repeat or vary their location choices.

The magnitude of location variability has important implications for location prediction. Most of location choice models in the transportation field are built with location choice information on a 'typical' day. In order to apply these models to predict one's location choices, individuals are assumed to repeat the same set of location choices over time. The likely existence of location variability raises the question that how individuals' location choices observed within

on this typical day can really be used to typify individuals' location choices in the long run. That is how much individuals are committed to the same set of locations over time. While current location choice models may produce satisfactory predictions in the case of low location variability, they could also yield rather distorted results for variable location choices.

Location variability characterizes individuals' activity location choice behavior which has been shown to be affected by time of day (15). The question that then arises is whether individuals' location variability depends on time of day. One rationale for expecting time-of-day dependence of location variability relates to the concept—level of fixity. Level of fixity is a concept on the extent to which activities are constrained in the time and space. Activities characterized with a higher level of fixity are more difficult to relocate and reschedule. Researchers have generally agreed that activities with the highest level of fixity are those compulsory ones, such as work or school (133). Since the various activities performed by an individual differ in their levels of fixity, it's conceivable that the location choices observed during time slots filled with activities with a higher level of fixity are less variable. The primary objective of this chapter is to examine the time-of-day dependence of location variability. Understanding time-of-day dependence of location variability will allow us not only to observe the variations in location variability with respect to time-of-day but also to quantify the impact of time-of-day on location variability.

In this chapter, first, individual temporal profile of location variability is constructed, i.e. individuals' location variability is measured for different time periods in a day. Sample means of location variability for different time periods are shown to be statistically different. In general, location choices are found to be the most variable in the afternoon and relatively stable in the

morning and evening. For each time period, significant heterogeneity in location variability is also observed.

In order to further explore this population heterogeneity, people are clustered into groups based on their temporal profiles of location variability. In this dissertation, model-based clustering is used to cluster relatively homogeneous temporal profiles together. Clustering based on the temporal files of location variability renders two distinct groups of people. The trend of location variability across the day is rather similar for the two groups, but one group of people consistently exhibit a higher level of variability. Further analysis shows that the between-group difference in level of location variability stems from the difference in the relative frequency of visits among locations instead of the total number of unique location visited.

Lastly, the magnitude of time-of-day dependence is quantified by introducing time period as independent variables in explaining location variability. All time related variables demonstrate significant explanatory power, although their magnitudes vary. In general, afternoon periods show a larger influence on location variability. Time variables collectively account for 36% of the variations in location variability. These facts confirm that time of day is an important factor that influences location variability.

The rest of this chapter is organized as follows. In Section 2, relevant literature on variability in spatial behavior and the concept of level-of-fixity is reviewed. Section 3 provides an introduction to the mobile phone data set, methods used to extract location information and to select a sample. In Section 4, methods used to measure location variability and cluster individuals based on their temporal profile of location variability are presented. Analysis results

can be found in Section 5. The chapter is concluded by a discussion of the implications of location variability on location choice models.

## 4.2 LITERATURE REVIEW

### 4.2.1 Location Variability

Typical approaches to evaluating location variability involve measuring the frequency of repeat visits to activity locations. Schönfelder (134) investigated people's activity locations and found the trips to the top two to four most visited locations (including home) accounted for more than 70% of all trips. Schönfelder and Axhausen (135) further showed that trips to the top 10 most visited activity locations accounted for 80% of all the trips and among them 40% were home-directed. Buliung et al. (106) found that, on average, people performed 72% of their activities at repeated locations and the remaining 28% were carried out at locations occurring only once during one week period. Most recently, Song et al. (56) examined the fraction of time a mobile phone user spent at his top-visited locations. The user was found to spend about 60% of his time at his top two locations. This number was reported to be 90% in a subsequent study (66). Lu et al. (65) elaborated these results by showing that these percentages could differ depending on the number of unique locations visited. On average, those who visited more than 10 locations spent approximately 75% of their time at the top two locations, while this percentage could be as high as 95% for those who only visited four distinct locations. These results consistently suggested that individuals' location choices presented a significant amount of repetition supplemented with some variability (134). However, no dedicated study on the impact of time-of-day on location variability is identified.

### 4.2.2 Level of Fixity

People travel between spatially separated locations to perform activities and the choices of activity locations are subject to the spatial constraints associated with these activities that can be represented by level of fixity. Level of fixity is a concept on the extent to which activities are constrained in space and time and rooted in time-geography (1). A key idea in time-geography is that individuals' activities and related travel are subject to spatial and temporal constraints. These constraints vary for different activities and can be illustrated with the space-time prism in Fig. 4.1.

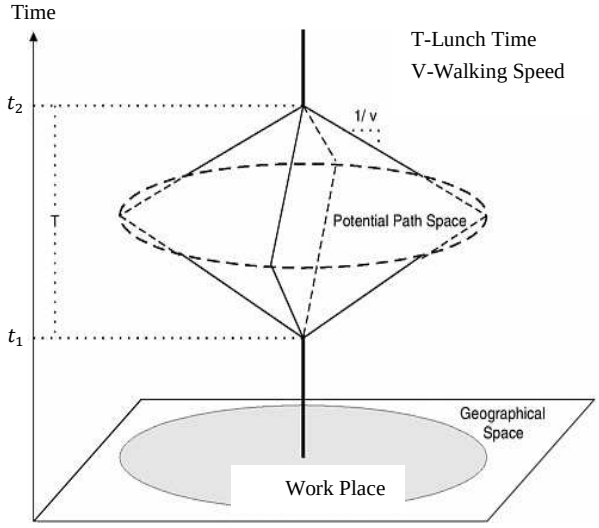


FIGURE 4.1 Illustrative example of spatio-temporal constraints on activities

Adapted from Wu and Miller (136)

The three-dimensional volume bounded by the space-time prism is called the *Potential Path Space* (PPS). In Fig. 4.1, the horizontal plane represents space and the vertical axis

represents time. This worker can't leave work place for lunch before $t_1$ and he is obliged to be back at work place at $t_2$, simply because his work commitments need to be carried out at his work place during specified working hours. On the contrary, lunch lends him more flexibility. Between $t_1$and $t_2$, the position he could possibly occupy in space and time can be delineated by the prism, assuming a constant walking speed $v$ represented by the slope of the edges of a prism. Within this prism, the place and time of lunch can be decided at his discretion. PPS is determined by this individual's time budget (i.e. $t_2 - t_1$), spatial constraints (work location that determines travel origin/destination within lunch time), and the travel speed $v$. Cullen and Godson (16) noted one of the features of this model as: there are two types of activities—fixed activities which are fixed in space and time (e.g. work) and unfixed activities which are relatively flexible (e.g. lunch).

Cullen and Godson (16) further elaborated this fixed-unfixed dichotomy of activities by arguing that people tended to attach a subjective level of fixity according to the extent to which an activity was constrained in time and space. Activities characterized with a high level of fixity are difficult to be relocated and rescheduled. Researchers have generally agreed that work and in-home-activities (e.g. sleeping) were of the highest level of fixity in an individual's daily activity-travel pattern (137, 138). 'Spur of the moment' activities (139, 140, 141) seem to have the lowest level of fixity and, oftentimes, are planned only a few minutes before they are actually performed. It is then conceivable that, within an individual's daily schedule, time periods filled with relatively fixed activities tend to lack location variability than the others.

## 4.3 METHODOLOGY

### 4.3.1 Entropy as a Measure of Location Variability

In this dissertation, location variability is measured by Shannon's entropy (142). Entropy is an established measure of variability in a random variable (143). Shannon's entropy $S$ for a discrete random variable $X$ can be mathematically written as:

$$S(X) = -\sum_n P(x_n) log_2 P(x_n), \quad [4.1]$$

where $P(x_n)$ is the probability of outcome $x_n$. $S(X)$ only takes on non-negative values and increases with greater variability. $S(X)$ measures variability in $X$ by capturing the number of unique values $X$ can take on and also the relative frequency of these values. Random variable with completely repetitive outcomes results in zero entropy, while a large number of outcomes with comparable probabilities of occurrence yield a larger entropy.

Let the location choices of individual $i$ during time period $k$ be represented by a random variable $J^k$. Location variability of individual $i$ during time period $k$—$S_i^k(J^k)$—can then be measured as

$$S_i^k(J^k) = -\sum_n P_i^k(j_n^k) log_2 P_i^k(j_n^k), \quad [4.2]$$

where $P_i^k(j_n^k)$ is the historical probability of individual $i$'s visiting location $j_n$ during time period $k$. Repetitive observations of individual $i$ being at a single location during $k$ would result in $S_i^k(J^k)$ being equal to zero, while regular visits to a large number of locations by individual $i$ yield a larger $S_i^k(J^k)$.

### 4.3.2 Temporal Profile of Location Variability

A temporal profile of location variability is a time-ordered sequence of location variability each computed for a specific time interval in a day. In order to construct temporal profile of location variability, a day is divided into $K$ equal-length time intervals. For each time interval $k$ ($1 \leq k \leq K$), the location variability of individual $i$ is computed as $S_i^k(J^k)$ and this individual's temporal profile of location variability can be represented as $\boldsymbol{S}_i = \{S_i^1, \dots, S_i^K\}$—an ordered sequence of entropy values each measuring the location variability during a time interval.

### 4.3.3 Model-Based Clustering

In this dissertation, model-based clustering is used to cluster temporal profiles of location variability. Conventional clustering algorithms (e.g. hierarchical clustering and k-means clustering) require a user-defined number of clusters. A more flexible algorithm that allows the number of clusters to be derived based on the data at hand is desired. Model-based clustering is a compelling alternative in achieving this objective (144).

Model-based clustering has seen its success in different applications recently (145, 146, 147). In model-based clustering, the observed data is assumed to be generated from a statistical model. Gaussian mixture model (GMM) is one of the most applied statistical models for model-based clustering analysis (148). A GMM is a weighted sum of a number of individual Gaussian distributions with unknown parameters. Each individual Gaussian distribution is a component of the mixture model and mathematically represents a cluster in clustering analysis. In determining the number of clusters, a user-specific upper bound of the number of components $M$ in a GMM, i.e. the maximum possible number of clusters underlying the data, is selected. For each number $m$ from 1 to $M$, the parameters of a GMM with $m$ component are estimated with expectation-

maximization algorithm (EM) (149) and the cluster membership is determined. Among all the estimated GMMs differing in number of components (1,…, $M$), the GMM with the highest Bayesian Information Criteria (BIC) is selected to be the best model generating the data. Consequently, the number of components in this best GMM is the resultant number of clusters. The resulting clustering membership from model-based clustering can be evaluated by an 'uncertainty' measure—the probability that a given observation doesn't belong to its assigned cluster. The smaller this probability, the more confident we are in the clustering results. Interested readers in model-based clustering are referred to Fraley and Raftery (144).

### 4.3.4 Linear Regression on Panel Data

In order to further quantify the dependence of location variability ($S$) on time of day, a linear regression model is built regressing location variability ($S$) on time of the day. Given the panel data nature of our data (with each individual having multiple entropy values calculated for different time periods), the model is specified as a linear regression with unobserved individual-specific effect. $K$ time periods will be entered as indicator variables ($INT1, \dots INT(K-1)$) in the model with the last time period as a reference.

The final model is specified as in equation [4.3].

$$S_i^k = \alpha + \beta_1 INT1 + \cdots + \beta_{K-1} INT(K-1) + u_i + \epsilon_{ik}, \quad [4.3]$$

where $i = 1, \dots, n$ is the individual index and $k = 1, \dots, K$ is the time period index. The idiosyncratic error $\epsilon_{ik}$ is assumed to be well-behaved i.i.d. white noise and $u_i$ is an individual-specific error term. Depending on the assumptions made for the relationship between $u_i$ and other regressors (correlated vs. uncorrelated), the model can be either estimated as fixed-effect model

and random-effect model. Fixed-effect model is chosen here, since the non-correlation between $u_i$ and other regressors can't be guaranteed.

## 4.4 DATA AND SAMPLE SELECTION

### 4.4.1 Data Overview

Please refer to Section 3.1 in this dissertation for a detailed description of the data.

### 4.4.2 Data Prepossessing

Please refer to Section 3.2 in this dissertation for a detailed description of the procedure.

### 4.4.3 Sample Selection

As discussed above, entropy relies on the knowledge of the probability of each location being visited. Since the probability of an individual choosing a particular location is not directly observable, it is approximated by the relative frequency of being visited in location history for the calculation of entropy. In general, observed frequency should become a better approximation as the study time period gets longer. This dissertation limits its scope to weekday travel. It is desirable to determine entropy based on one's travel record on all 43 weekdays throughout the whole study period, which is referred as real entropy ($S_{real}$) in this section. However, about 47% of our sample is found to be observed for less than 5 days. Moreover, as sightings don't spread uniformly within a day, location information during certain time intervals can be missing. As a matter of fact, not a single individual in our sample has location choice information for each hour on all days. In order to select a sample of reasonable size to derive statistics, a concept of operational entropy ($S$) is proposed here to resolve this problem.

Operational entropy can be calculated based on location information from those days when an individual is observed. Operational entropy serves as a reasonable approximation of the real entropy only if the number of days with missing location information were within a certain threshold. In other words, if an individual was observed for $n$ days and $43 - n^* \leq n \leq 43$, where $n^*$ is the threshold to be determined (see next paragraph), the operational entropy calculated based on the location choice information on these $n$ days is considered to be a reasonable representation of the real entropy.

In order to determine $n^*$ the following experiment was conducted. First, a subsample of 10,533 individuals who were observed for all weekdays during the study period are selected so that their real entropy ($S_{real}$) is known. Next, $m$ $(1 \leq m \leq 42)$ days are randomly removed from these individuals' records to simulate the circumstances that the location choice information is only available on $n = 43 - m$ days. Finally, operational entropy ($S$) is calculated based on available information on these $n$ days. In order to evaluate the deviation of operational entropy ($S$) from real entropy ($S_{real}$), ratios of $S$ to $S_{real}$ with respect to the number of days removed $m$ is examined in Fig 4.2.

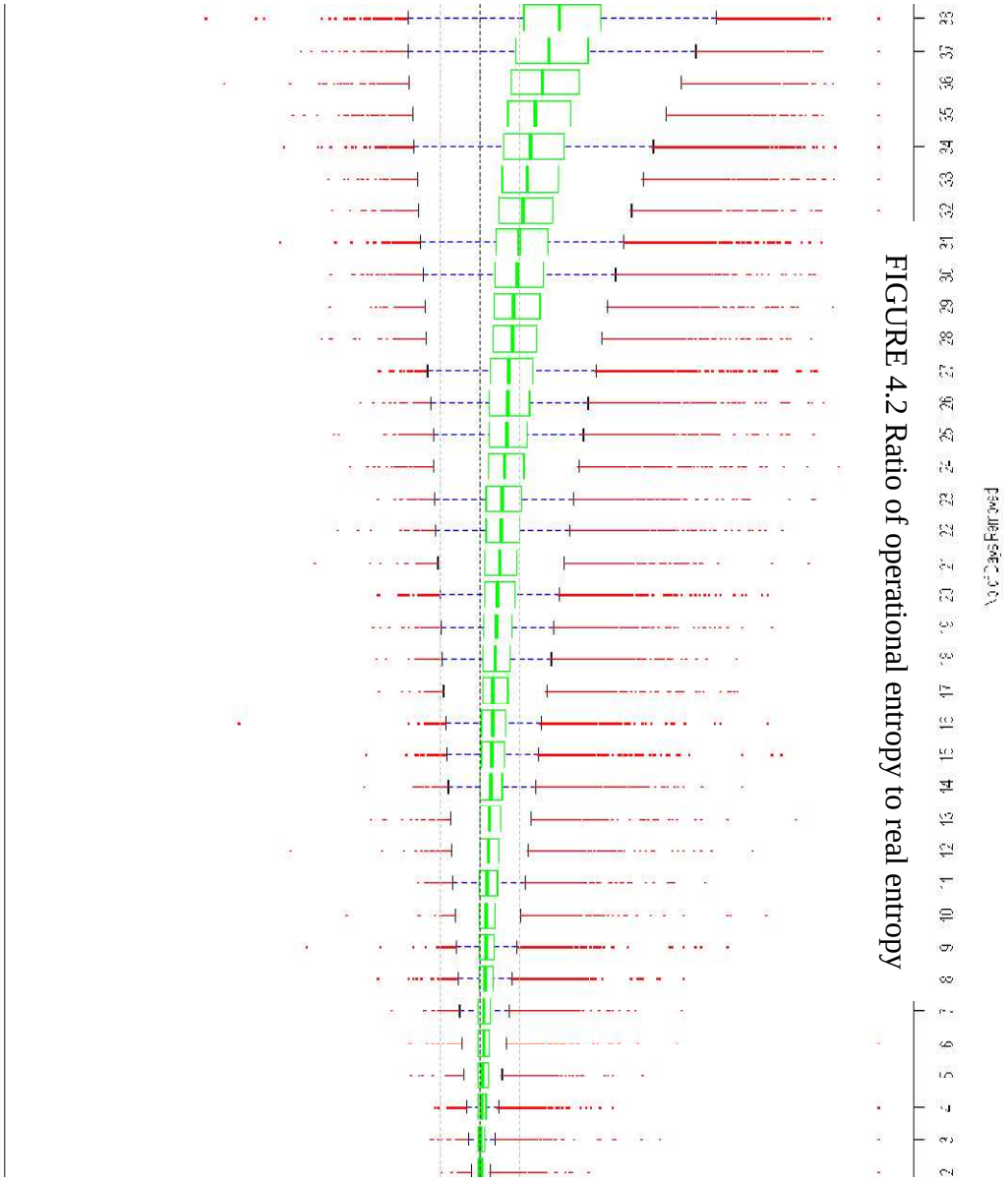

FIGURE 4.2 Ratio of operational entropy to real entropy

Fig. 4.2 shows the boxplot for the ratio of operational entropy $S$ to the real entropy $S_{real}$. Green boxes cover the range between the first quartile and the third quartile with the median value clearly marked in the middle. The blue whiskers extend to 1.5 times the range between the first quartile and the third quartile and the red dots are outliers. A ratio of value one indicates that the operational entropy $S$ is equal to the real entropy $S_{real}$. When only 1-day data was removed, for the majority of the sample, the real entropy $S_{real}$ is very well approximated by the operational

entropy $S$: the median of the ratio $S/S_{real}$ is one and those values between first quartile and third quartile are extremely close to one. In addition, only a few operational entropy values fall out of the range between 0.9 and 1.1 times of the real entropy. As more days are removed, the medians deviate further from one and there is more variations in the ratio as shown by the increasing box size. More and more operational entropy values are now beyond the [0.9 $S_{real}$ , 1.1 $S_{real}$ ] window. In general, operational entropy deviates farther from real entropy with larger number of days removed.

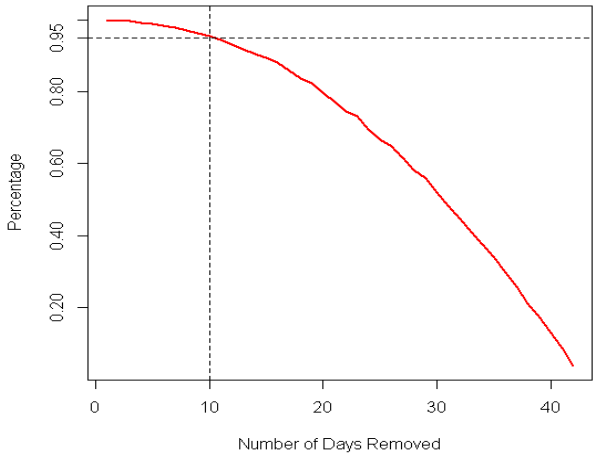


FIGURE 4.3 Percentage of operational entropy within ten percent range at real entropy

In this dissertation, entropy falling in a window [0.9$S_{real}$ ,1.1$S_{real}$ ] is considered as a reasonable approximation of real entropy. This selection is heuristic. A different window size could have been selected. Then the percentage of the sample having an operational entropy ($S$) as a reasonable approximation of the real entropy ($S_{real}$) is identified for different numbers of days removed ($m$). The results are shown in Fig. 4.3. Fig. 4.3 shows that the percentage of operational entropy within the window decreases with more days removed. With 1 day removed,

99% of the sample have an operational entropy falls in the window, while with 42 days removed, this percentage drops to 4%.

It is desirable that a higher percentage of the selected sample has an operational entropy falling in the window $[0.9S_{real}, 1.1S_{real}]$, i.e. an unbiased representation of the real entropy. Yet, the higher percentage, the fewer days removed. That is, higher percentage requires the selected sample has fewer days with unknown location information and thus results in a smaller sample size. In this dissertation, it is decided that a sample with 95% of the sample having operational entropy within the range would suffice. The sample is, therefore, selected in a way such that individuals in the sample have no more than 10 days with unknown location choice information (because the maximum number of days associated with more than 95% of sample having operational entropy within $[0.9S_{real}, 1.1S_{real}]$ in Fig. 4.3 is 10 days).

### 4.4.4 Number of Time Periods

Since sighting generation depends on phone activity and, for majority users, no phone activity is performed during nighttime, the total number of nighttime sightings is limited (less than 7% of all sightings were generated during 12 am to 6 am). Based on this observation, only individuals' location choices during daytime are analyzed (6 a.m. to 12 a.m. the next day). Determining the number of time periods requires the selection of an appropriate interval length. Theoretically, short time intervals are desirable in order to capture variations of location variability over time. Yet, too short a time interval would lead to a large number of interval with missing location information and thus too small a sample size.

For each time interval $k$ $(k = 1, \dots, K)$ on each day, it can be determined that whether an individual has location choice information within this interval. The number of days with missing

location choice information during time interval $k$ is $m_k$. According to the threshold determined in the last section, an individual is only selected if all $m_k's$ are no larger than 10. This approach would produce samples of different sizes depending on the number of time intervals $K$ in a day. Experiments with $K = 3,6\ and\ 9$ result in sample sizes of 12,483, 2,492 and 774 individuals, respectively. A comparison between these samples and original sample is shown in Table 4.1 in terms of the mean of average daily number of sightings and the mean of average time intervals.

Table 4.1 Sample Comparison

| | Average daily number of sightings | Average time interval (min) |
|---|---|---|
| | Mean | Mean |
| Original Sample | 32.90 | 18.62 |
| Sample | | |
| 3-interval | 90.13 | 13.56 |
| 6-interval | 155.55 | 8.46 |
| 9-interval | 195.38 | 8.00 |

Our samples generated a much larger number of sightings on a typical day compared to the original sample and had much shorter time intervals between consecutive sightings. In general, our samples represent those who had more phone activities in the population. These attributes become more salient when smaller time intervals are considered. Therefore, caution should be exercised when extrapolating the results to the population. This issue will be discussed in more detail in the last section of this chapter.

## 4.5 RESULTS

### 4.5.1 Sample Distribution of Location Variability

The distributions of entropy for different time interval divisions (i.e. three, six and nine intervals) are shown in Fig. 4.4.

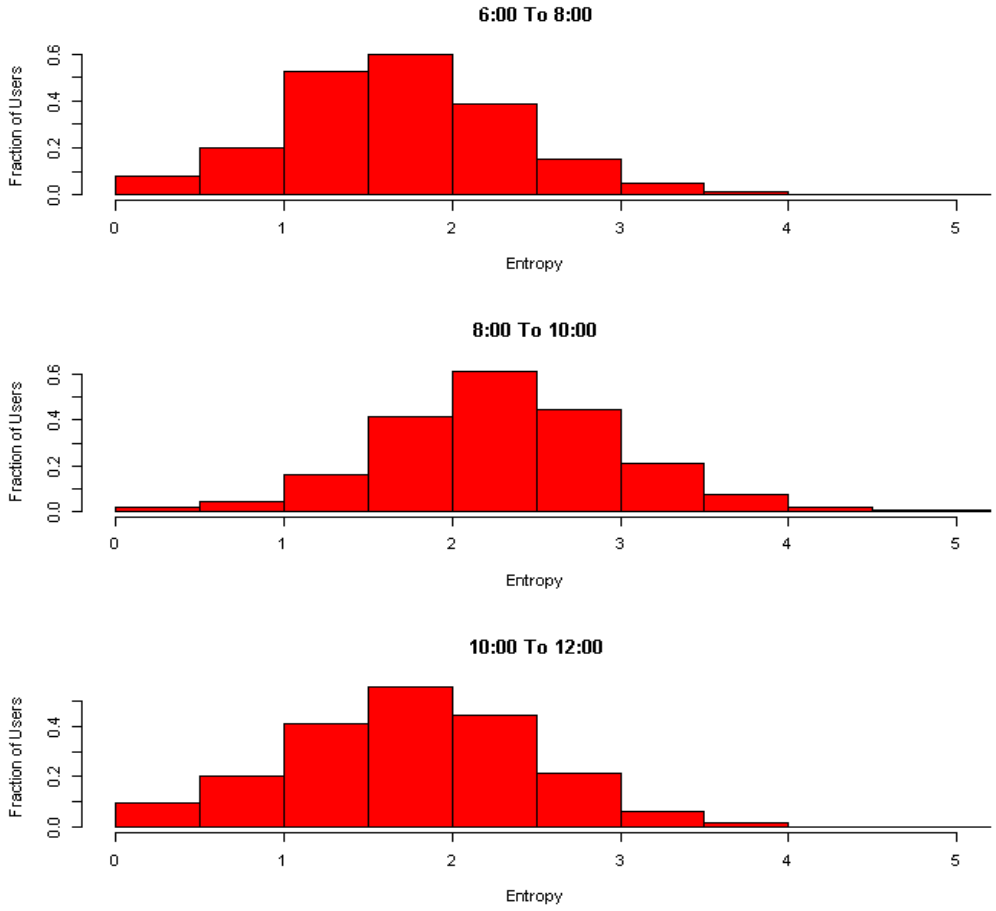


FIGURE 4.4a Distributions of entropy for three intervals in a day

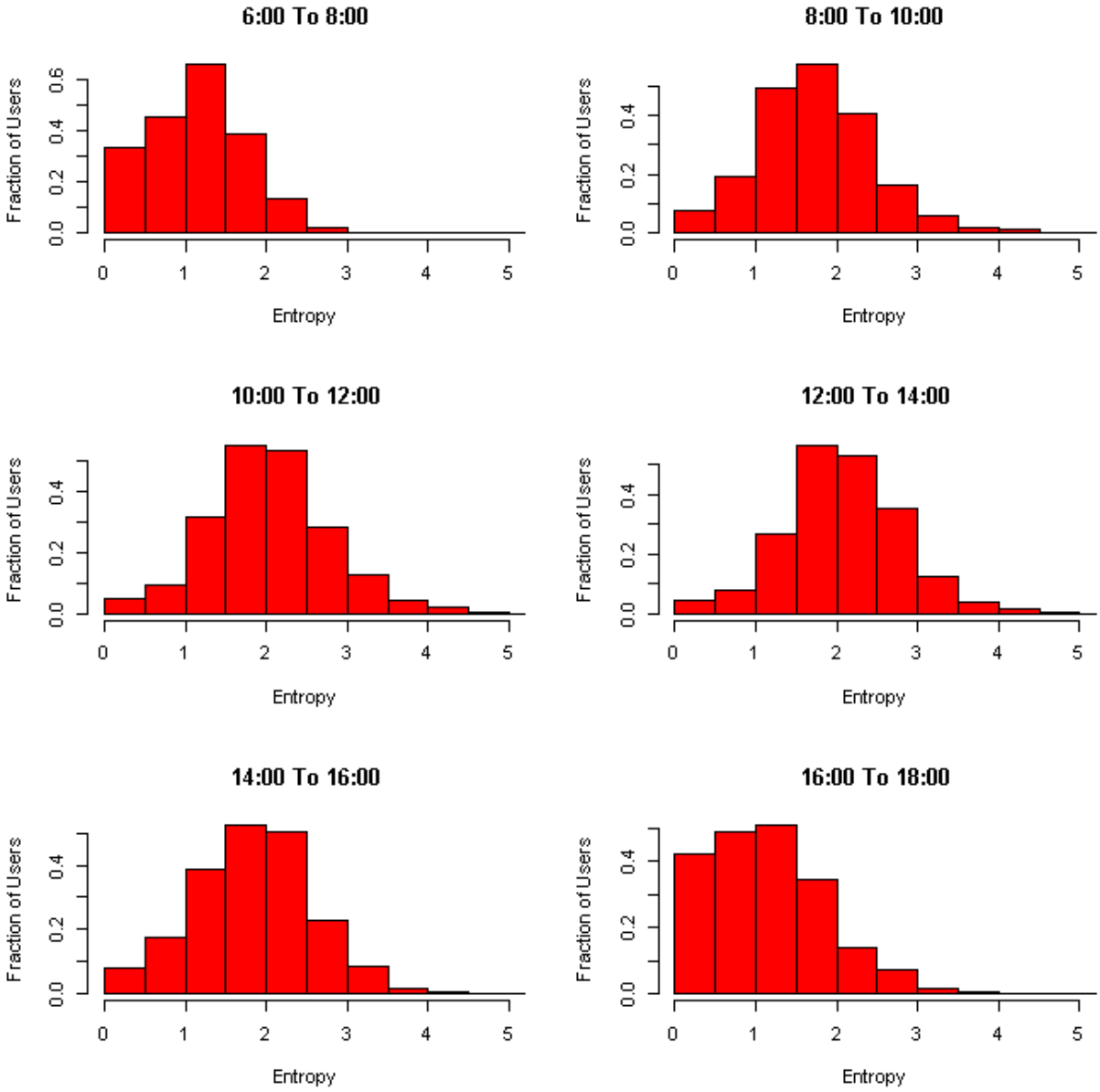


FIGURE 4.4b Distributions of entropy for six intervals in a day

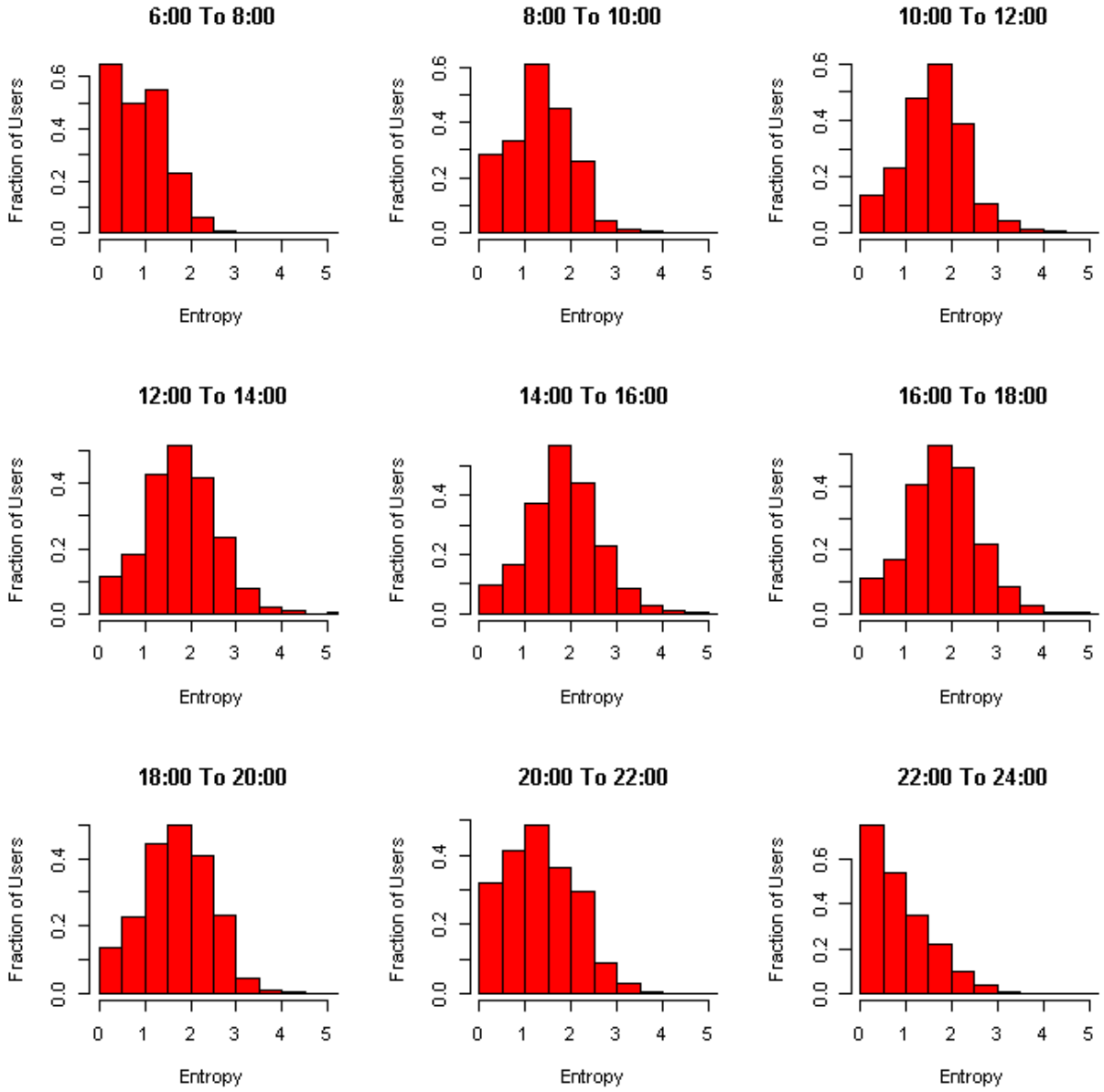


FIGURE 4.4c Distributions of entropy for nine intervals in a day

In general, individuals' entropy values fluctuate within the range from 0 to 4 and the distributions between different time intervals are visually distinct. As the time intervals become shorter, more subtle variations in the entropy distribution with respect to time are revealed. In Fig. 4.4c, for the time period from 6 a.m. to 12 p.m., the distribution shifts to the right with time:

most individuals have a low entropy value smaller than 2 from 6 a.m. to 8 a.m.; as time passes, entropy increases; when it approaches noontime, the distribution peaks at around 2. For the time period from 12 p.m. to 6 p.m., the distribution doesn't seem to change significantly from one time interval to another: all the distributions are bell-shaped with a single peak around 2, though a slight increase in entropy with time can be spotted. Contrary to that observed from 6 a.m. to 12 p.m., peaks of distributions gradually shift to the left during the time period from 6 p.m. to 12 a.m. next day: though the distribution during 6 p.m. to 8 p.m. is still comparable to those in the afternoon period, for most of the people, entropy during 8 p.m. to 10 p.m. shows a significant decrease and continues to decrease as it approaches midnight.

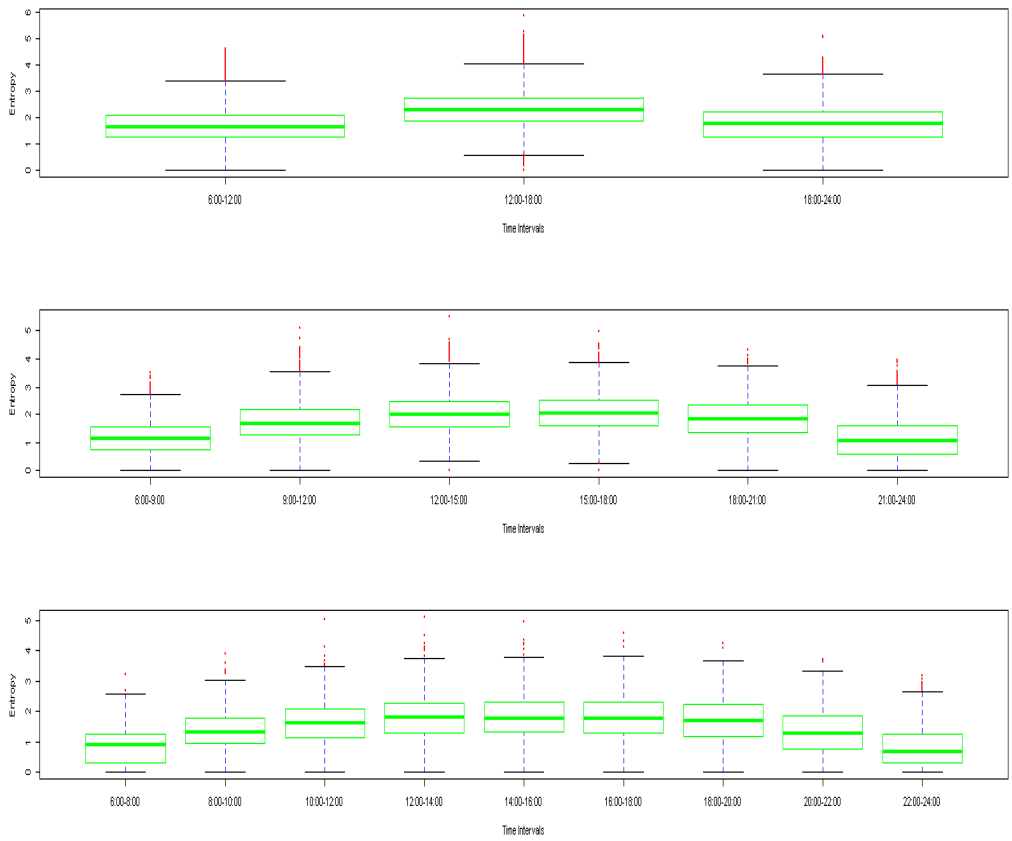


FIGURE 4.5 Boxplot of entropy by time intervals

Fig. 4.5 illustrates the temporal profile of location variability, i.e. the distribution of entropy by time intervals. Green boxes cover the range between the first quartile and the third quartile with the median value marked in the middle. The blue whiskers extend to 1.5 times the range between the first quartile and the third quartile and the red dots are outliers. Despite the number of intervals specified, the temporal profiles of location variability revealed are highly consistent: entropy increases in the morning, reaches its maximum in the afternoon, and decreases in the evening. This finding suggests that individuals are most likely to repeat their previous activity location choices in the morning and tend to be the most flexible in activity location choices in the afternoon. These results also support the hypothesis that there exist time-of-day dependence in location variability.

### 4.5.2 Clustering Individual Temporal Profile

The temporal profile of location variability $\boldsymbol{S}_i$ observed at aggregate level could have two alternative explanations: 1) individuals have similar temporal profiles characterized with highest variability occurring in the afternoon and relatively low variability in the morning and evening; 2) the observed temporal profile can be comprised of distinct profiles. Therefore, we further apply model-based clustering to search for relatively homogeneous temporal profiles $\boldsymbol{S}_i$. For the best temporal resolution, the following are the results derived based on the nine-interval division and a sample including 774 individuals.

The maximum possible number of clusters underlying the data is selected to be 9 in this dissertation. Consequently, for each number $m$ from 1 to 9, the parameters of a GMM with $m$ component are estimated. Bayesian Information Criteria (BIC) of all these estimated GMMs differing in number of components (1,…, 9) are -9393, -9359, -9395, -9438, -9391, -9417, -9418, -9415 and -9430, respectively. Since the GMM with the highest Bayesian Information Criteria

(BIC) is considered to be the best model generating the data, model-based clustering yields 2 clusters of temporal profiles (with BIC equal to -9359), i.e. 2 individual Gaussian distributions as component of the mixture model.

The first group comprises 173 individuals and the second group includes 601 individuals. The resulted clustering membership resulted from model-based clustering can be evaluated by an 'uncertainty' measure—the probability that a given observation doesn't belong to its current assigned cluster. Quartiles of the 'uncertainty'—the probability that a temporal profile doesn't belong to its current assigned cluster are 9.909492e-04 (first quartile), 9.056369e-03 (second quartile) and 9.051605e-02 (third quartile), respectively. Small values of these numbers indicate that the majority of temporal profiles are well classified. Changes in mean entropy over time for both groups are displayed in Fig. 4.6.

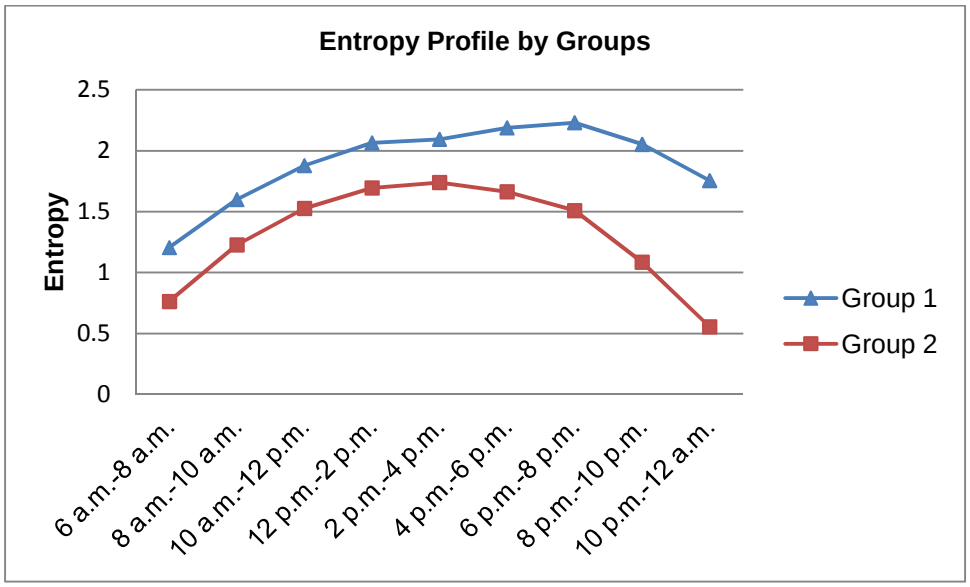


FIGURE 4.6 Temporal profiles of location variability by group

The overall trend of these two temporal profiles is very similar and resonates with what is observed at aggregate level (Fig. 4.5). Individuals exhibit more repetitious location choices in the

morning and evening and a relatively high level of variability in the afternoon. It is also evident that individuals in group 1 constantly exhibits higher entropy values, i.e. level of variability, compared to those in group 2. As mentioned above, entropy as a measure of location variability is determined by both the total number of location visited and the visit frequency to each location. These two quantities are further examined to identify potential cause of the difference in location variability between groups. First, a t-test shows no significant difference in the total number of location visited. Both groups visited an average 18 unique locations during these two months. In order to examine the visit frequency to each location, locations visited by individuals are ranked by their visit frequency and the visit frequency ratio of top *L* visited locations (i.e. the number of visits to top *L* visited locations divided by the total number of visits to all locations) is plotted for each group (Fig. 4.7).

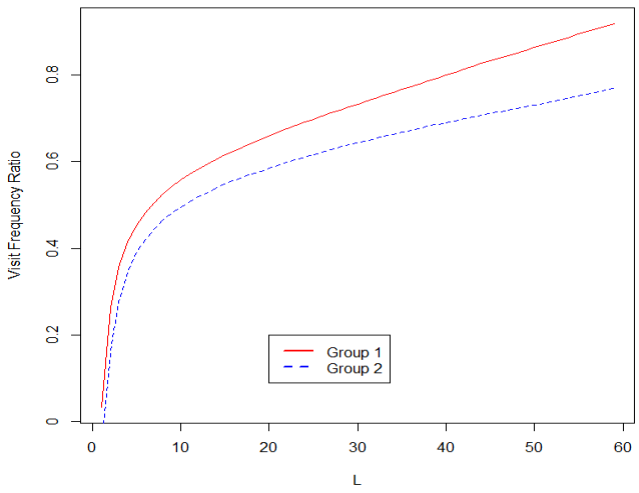


FIGURE 4.7 Visit frequency ratio of top *L* visited locations by group

It is clear that individuals in group 1 generally spend more time at a few locations that are frequented, while individuals in group 2 spend their time more evenly at other less regularly

visited locations. So the difference in location variability between these two groups of people is more likely to stem from the way people distribute their time among different locations, rather than the number of locations they explored.

### 4.5.3 Regression Model

The graph above (Fig. 4.6) illustrates that, despite population heterogeneity, the effect of time-of-day on entropy is very similar between groups. In this section, a linear regression model is specified with entropy explained by time periods as indicator variables to further quantify this time-of-day effect on location variability. Estimation results of the model are presented in Table 4.2. The time period used as a reference is the two-hour time period from 10 p.m. to 12 a.m. the next day. The results show that, except for the first time period (6 a.m. to 8 a.m.), other time periods are associated with a level of location variability that is significantly higher than that during the reference period. Among them, location variability in afternoon time periods is the highest. This finding is consistent with what is observed in the descriptive analysis above. Time variables collectively account for about 36% of total variations in location variability. This indicates time-of-day is an important factor that influences location variability.

Table 4.2 Estimation Results

| | Estimate | Std. Error | t-value | p-value |
|---|---|---|---|---|
| (Intercept) | 0.52* | 0.05 | 10.29 | <0.000 |
| $INT1$ (6 a.m-8 a.m.) | 0.03 | 0.03 | 1.19 | 0.23378 |
| $INT2$ (8 a.m-10 a.m.) | 0.48* | 0.03 | 18.79 | <0.000 |
| $INT3$ (10 a.m-12 p.m.) | 0.77* | 0.03 | 30.39 | <0.000 |
| $INT4$ (12 p.m-2 p.m.) | 0.94* | 0.03 | 37.19 | <0.000 |
| $INT5$ (2 p.m-4 p.m.) | 0.99* | 0.03 | 38.81 | <0.000 |
| $INT6$ (4 p.m-6 p.m.) | 0.95* | 0.03 | 37.41 | <0.000 |
| $INT7$ (6 p.m-8 p.m.) | 0.84* | 0.03 | 33.15 | <0.000 |
| $INT8$ (8 p.m-10 p.m.) | 0.48* | 0.03 | 18.73 | <0.000 |
| R-squared | 0.36 | | | |
| F-statistic | 347.997 | | | |
| p-value | <0.000 | | | |

Note: * indicates a significance level at 0.05.

## 4.6 CONCLUSIONS AND DISCUSSIONS

In this chapter, the time-of-day dependence of location variability is investigated. Individuals are found to be more likely to vary their location choices in the afternoon periods than in the morning and evening. This result is consistent with previous findings in activities scheduling behavior. Joh et al. (150) studied modifications (e.g. change of start and end time, location and accompanying person) made by individuals on pre-planned activities and showed that those activities scheduled in mornings were the least likely to be modified, followed by those scheduled in evenings and then those in the afternoons. This high likelihood of modification on

activities scheduled in afternoon will definitely interrupt some routine in location choice and contribute to variability.

Yet, it is not the time-of-day dependence of location variability that is surprising. Rather, it is how significantly time-of-day effect shapes location choice behavior. Despite population heterogeneity, individuals exhibit similar temporal profile of location variability and time-of-day takes account for approximately 36% of the total variations in location variability. This finding confirms the importance of time as a factor that influences human spatial behavior. The importance of time in human spatial behavior can be traced back to Hägerstrand's time-geography (1). Hägerstrand described three classes of constraints on spatial activities: capability constraints—limitations on the activity of an individual "because of his biological structure and/or the tools he can command"; coupling constraints—limitations that "define where, when, and for how long, the individual has to join other individuals, tools, and materials in order to produce, consume, and transact"; authority constraints refer to "domain" or "a time-space entity within which things and events are under the control of a given individual or a given group."

The first class of constraints suggests that human beings are regulated by biological clocks that prefer conducting certain activities on a regular basis (e.g. sleeping during night time every day). Consider an individual who must get back home before midnight. As time proceeds to later in a day, the available time for travel is gradually reduces. Hence, this person is more likely to visit locations in the proximity of his home in order to get home on time, which leaves less flexibility in location choice. This would probably explain the low location variability during the evening period. On the other hand, the relatively low location variability in the morning and early afternoon is more likely to be explained by individuals' needs to meet the coupling constraints, i.e. the necessity of synchronizing one's behavior with others. Modern

society is regulated by clock time (e.g. work hours from 9 a.m. to 5 p.m.). Since people need to abide by these regulations in order to interact with others (2, 151), they are left with limited choices of location to perform activities that require predictable presence by others. For instance, a worker is expected to be at his workplace during work hours.

Studying the time dependence of location variability, i.e. how people repeat or vary their location choices depending on time, has important implications on predicting people's location choice behaviors. Recently, there has been much interest in exploiting the repetitious nature of human travel in location prediction (65, 152). Underlying these location prediction models is the assumption that people tend to repeat to the same set of locations they visited before (121, 153). Though the existence of repetitious visits is evident (56, 57), little is known about the temporal characteristics of these repeated visits, i.e. when these visits are most likely to be repeated. In fact, the intensity of repeated visits greatly depends on its temporal context. In Song et al. (56), the authors noted that the probability of an individual returning to the most-visited activity location during a specific hour differed across a day. During nighttime, when most people tend to be at home, this probability peaks at 0.9, while during some transition periods during a day (e.g. lunch time), e.g. from noon to 1 p.m., this probability hits the minima at only a little over 0.5.

A few meaningful attempts have been made to factor in the temporal information in making location predictions (18, 19) and have obtained promising results. Taking account of temporal dimension in modeling location choice behavior will yield a deeper understanding of human spatial behavior and advance the development of location choice modeling. This dissertation contributes to the research by demonstrating significant effects of time on location variability.

# CHAPTER 5

# MORE EFFICIENT LOCATION PREDICTIONS

## 5.1 INTRODUCTION

Location prediction provides the basis for a wide range of applications, including enhancing the performance of cellular wireless communication network (154, 155, 156, 157), supporting effective transportation management (119, 158) and enabling more intelligent location-based services (159, 160). Over the past decades or so, a myriad of location predictors have been developed based on various modelling techniques (161), such as Markov chain-based predictors (3, 162, 163, 164), data mining-based predictors (21, 159, 165) and neural network-based predictors (70, 166, 167, 168), to name a few.

Many of the above predictors rely heavily on individuals' past location patterns as input (169) based on the high degree of temporal and spatial regularity (56, 57) in human mobility. Explicitly or implicitly, these predictors are developed based on the understanding that there are patterns in individuals' past movements, i.e. recurring components, which they tend to repeat in the future. Further developments are expected to be made to this class of predictors as the advent of location-aware devices has greatly reduced the effort involved in location history collection.

In general, a longer location history results in more accurate predictions, simply because recurring components, defined by its repetitious occurrence in individuals' location history, become more distinguishable from random components over time (170). Yet, it is also well recognized that storing and processing a long location history for a large number of people is impractical in many aspects. First, it poses a challenge to data storage. A large amount of memory needs to be reserved for this type of data, especially with the increasing popularity of

location-aware applications. Location data can now stream in at an unprecedented rate, data size can easily become unmanageable. Second, processing this sheer amount of data can be time consuming and render poor computational efficiency. Computational efficiency in the location prediction context describes the extent to which resources, including time and storage, are well used to make location predictions. Location prediction characterized with higher computational efficiency is fast and space saving. Computational efficiency is becoming more important as there is an increased interest in making location prediction with the limited memory of mobile devices (71, 171).

Different strategies have been proposed to improve computational efficiency in location prediction in existing studies. Some strive to develop more efficient location prediction algorithms (171) and others suggest to implement the more time-consuming and computationally-costly predictor training phase in an offline manner (20, 172). Yet, little attention has been directed to the possibility of reducing the size of input information, i.e. the length of location history in the case of location history as input. Removing partial history is tied to the concern about loss of valuable information and thus less accurate predictions, as researchers have always had the conviction that less information would lead to a sacrifice in accuracy (69). This is, however, not necessarily the case if that part of history contains little information.

The amount of information carried by each observed location in location history varies among individuals. If an individual's location choice is thought as a random variable, the average amount of information contained in an observed location can be quantified by entropy. Entropy (referring to Shannon's entropy here) is an established measure of the expected amount of information one needs to determine the value of a random variable, or equivalently, the

uncertainty in random variable (142). A random variable with high entropy contains more uncertainty and each observed outcome for this variable contains more information. Therefore, it is hypothesized that an observed location in a trajectory of low entropy contains limited information and thus removing some observations may not render a significant effect on prediction accuracy. Intuitively, for an individual who tends to repeat the same set of location choices every day and thus has little uncertainty, his location history on the previous day is likely sufficient to make an accurate prediction of his location choices on the current day. This could allow us to improve the efficiency of location prediction without compromising prediction accuracy.

The primary objective of this chapter is to examine the correlation between the length of location history and prediction accuracy for subpopulations differing in the level of uncertainty in their trajectories. The level of uncertainty in the trajectories of 3,568 mobile phone users (see Section 5.3.4 for sample selection) is measured with entropy, which is used to categorize them into four groups. For these four groups of people, changes in prediction accuracy with respect to increased history length are compared between groups. It is found that we can predict correctly the next location approximately 70% of the time for the most regular subgroup, given only 20 of their recent (not necessarily unique) locations. On the other hand, for those whose movements are the most irregular, prediction accuracy level only reaches a little over 50% given a hundred of historical locations.

The location is predicted by order-1 Markov predictor. Markov predictor represents a major family of the location prediction models. It has seen its success in many applications. More importantly, in a recent comparison study of multiple location predictors (173), Markov predictor was the best in terms of both prediction accuracy and efficiency. This dissertation also

serves as a valuable addition to the existing literature experimenting on the use of Markov predictors to cellular network.

The rest of this chapter is organized as follows. In Section 2, relevant literature is reviewed. Section 3 provides an overview of our data, data pre-processing procedure and sample selection procedure. Analysis results are presented in Section 4. The chapter is completed by a conclusion and some discussions in Section 5.

## 5.2 LITERATURE REVIEW

### 5.2.1 Entropy as a Measure of Uncertainty

A number of recent studies have measured the uncertainty in one's trajectory with entropy $S$. A user's trajectory with $S = 2$ can be interpreted as the uncertainty in this user's whereabouts is $2^S = 2^2 = 4$ locations, that is, this user can be found in any of $2^S = 2^2 = 4$ locations. Song et al. (56) assigned three entropy measures to each individual's trajectory:

1) the random entropy which assumes each location is visited with equal probability,

$$S_i^{rand} \equiv log_2 N_i, \quad [5.1]$$

where $N_i$ is the number of unique locations visited by user $i$;

2) the temporal-uncorrelated entropy

$$S_i^{unc} \equiv -\sum_{j=1}^{N_i} p_i(j) log_2 p_i(j), \quad [5.2]$$

where $p_i(j)$ is the historical probability that location $j$ was visited by the user $i$;

3) the actual entropy,

$$S_i = -\sum_{T_i' \subset T_i} p(T_i') log_2\, p(T_i'), \quad [5.3]$$

where $p(T_i')$ is the is the probability of finding a particular time-ordered subsequence $T_i'$ in the user $i's$ trajectory $T_i = \{X_1, \dots X_L\}$—an time-ordered sequence of locations user $i$ visited.

While the random entropy assumes each location is visited with an equal probability, the temporal-uncorrelated entropy takes account for the relative frequency of visiting all the locations, but not the temporal order in which the locations are visited. The real entropy captures not only on the frequency of the visitation, but also the order in which the locations are visited. They showed, for 50,000 individuals, while the random entropy peaked around six, the real entropy peaked at 0.8, which indicated the real uncertainty in a typical individual's location is only $2^{0.8} = 1.74$ locations. Yet, despite the entropy measure used, a great amount of heterogeneity was observed. The real entropy had a range from zero to more than 2. Lu et al. (65) applied the same set of entropy measures under a different setting. They arrived at similar conclusions by showing the real entropy of a typical user's was as low as 0.71, which leads to an uncertainty level as low as $2^{0.71} = 1.64$ locations. It's worth to note that the entropy range was also very close to that in (56). These results were shown to be rather robust even under extreme conditions. Lu et al. (66) investigated the mobility patterns of users affected by the earthquake in Haiti in 2010 and identified a slightly higher, even though still rather low, uncertainty level in human trajectories. They reported an uncertainty of $2^S = 2^{1.5} = 2.8$ locations with the real entropy peaks at 1.5. Heterogeneity in entropy was also evident with real entropy ranging from a little over 0 to approximately 4.

### 5.2.2 Location Prediction in Transportation Field

Location prediction has been addressed as a choice problem modeled by discrete choice models in the transportation field (174). The discrete choice model framework for location prediction can be described as following: each individual has a set of available locations—the choice set; each alternative location is assigned a utility (i.e. a satisfaction index) by the individual depending on the attributes of the alternative and the individual, the context, the values the individual attaches to these attributes, and a random utility part capturing the unobserved factors. In other words, the utility of each location is a random variable whose value is based on two parts: the systematic utility that can be observed and the random unobserved utility; the location with the highest utility is then selected as the predicted location. Depending on the distribution specified for the random utility term, a myriad of discrete choice models have been developed. Since introduced by McFadden (175), the multinomial logit model (MNL) remains a popular model in location prediction (176, 177) due to its simple closed-form. In MNL, the random utility term is assumed to be independent and identically Gumbel distributed. This specification brings simplicity in estimation. However, this specification requires independence among alternatives (178). Therefore, MNL can lead to erroneous predictions when there are unobserved similarities between alternatives, as is often the case in location choice. Unobserved similarities among alternative locations can stem from spatial adjacency (176). The reasons to expect similarities among adjacent locations are: 1) they provide similar activity opportunities due to the continuity of space; 2) they are aligned along the same transportation corridor with similar accessibility. Additionally, locations sharing common attributes (e.g. same land use type) are also considered similar.

Earlier amendments to MNL in order to capture this unobserved similarity among alternative locations include competing destination (CD) model (78, 179, 180, 181, 182, 183,

184, 185) which argues that the discrete choice model is a model originally developed in a non-spatial scenario and thus reflects a different choice-making process from that of spatial choice (78). Specifically, individuals do not simultaneously evaluate all location alternatives as assumed in a traditional MNL. In contrast, they employ a hierarchical decision making process in which a cluster of locations is chosen prior to the selection of a single location from this cluster. In other words, some alternative locations, in those unselected clusters are not evaluated. CD models were developed to model this hierarchy decision making process. Utilities of alternative locations are weighted by the likelihood that an alternative actually falls into a selected cluster and subsequently evaluated. This likelihood specification allows the modeler to avoid making deterministic judgments regarding the cluster membership of location alternative. This type of model differs from nested logit model (an advanced form of MNL supporting hierarchical selection) by embracing the notion that the composition of clusters of location alternatives perceived by individuals is unknown to the modeler (186, 187, 188). Although taking on a form similar to the MNL, with the use of likelihood function, CD models relax the IIA property of MNL and recognize that similarities among alternative locations (78, 177). Though intended to distinguish itself from aspatial discrete choice model, the CD model was recognized as a restrictive case of discrete choice model accounting for individual's limited knowledge on choice set (189).

In fact, many enhancements to discrete choice model have been made in recent years allowing researchers to account for the unobserved similarities among alternative locations (189). Among them, models belonging to the Generalized Extreme Value (GEV) family are most popular. All models in the GEV family nest MNL. In a GEV model, the random utility of each alternative location is decomposed into a common component shared among all the alternatives

in one nest and a component unique to each alternative. This decomposition permits interdependence of random utility among alternatives in one nest and thus it's possible to capture unobserved similarity among alternative locations. Several models belonging to the Generalized Extreme Value (GEV) family have been successfully applied in a variety of transportation-related choice problems (e.g. mode choice, route choice and departure time choice), including ordered GEV (190), crossed-nested logit (191, 192, 193, 194), MNL-ordered GEV (195), paired combinatorial logit (196), Generalized Nested Logit (197), generation logit (198) and distance-based GEV (199).

However, limitations remain in applying GEV models to explain the location choices of individuals. One limitation relates to the specification of systematic utility. GEV models, same as the MNL, specify the systematic utility as a deterministic weighted function of a set of observed attributes and individual characteristics with constant coefficients. This practice assumes homogeneous values that individuals attach to the attributes of alternatives. In other words, random effects due to individual differences cannot be accounted for. This problem has been paid little attention in studies on location choice with a few exceptions (178, 200). One common approach to address this problem is to estimate a mixed logit model, which allows the analyst to capture the random effect. The number of studies employing mixed logit model has been rapidly growing in recent years (201). Yet, a caveat remains: great care must be taken to ensure identification of these models (202). As noted in Walker (201), a large number of simulation draws and multiple model estimation runs were required to verify the identification and parameter stability of a mixed logit model. There are always trade-offs between the mixed logit model and other closed-form models to be made (mixed logit, probit, and GEV) regarding computational and performance issues. A parallel recommendation was made by Bhat (203):

researchers must always explore alternative closed-form models before turning to open-form models.

Another limitation that is conventionally overlooked in previous discrete choice models of location prediction is time-dependent choice. The location chosen at a current occasion is assumed to be independent of the choices from previous occasions. However, studies have shown that individuals present a considerate amount of variety-seeking and/or loyalty in location choices (56, 57, 204). This observation necessitates the integration of historical location choices into discrete choice models. Few studies have explicitly addressed this issue. Keperman et al. (204) added in a separate term representative of the impacts of the last location choice in the utilities of location alternatives at current occasion. Sivakumar and Bhat (205) devised the utility function of each location alternative with an addition of a term indicating whether the last location choice is the same as a current alternative. In two papers concerning individuals' residential location choices, Chen et al. (127) and Chen and Lin (206) specified the utility function of each residential location alternative to reflect the influence of the attributes of prior residential locations. However, these studies were significantly limited by the available information in the data sets regarding individuals' past location choices. Fortunately, the lack of history-dependence in location choice models has been addressed by many recently developed location predictors, including the Markov predictor discussed in the next section.

It is also worthy to note that discrete choice models are developed based on utility maximization theory and inherently behavioral models. The development of discrete choice model reflects researchers' understanding of the underlying mechanism of location choice behavior. On the other hand, the Markov predictors yet to be discussed in the next section rely solely on the spatio-temporal information contained in one's trajectory. Though Markov

predictors well serve location prediction purposes, these models don't necessarily reveal any insights into complex behavior underpinnings.

### 5.2.3 Order-$k$ Markov Predictor

Order-$k$ Markov predictor belongs to one of the most widely applied predictor families—the Markov chain family (207). Order-$k$ Markov predictor assumes that an individual's choice of the next location only depends on its $k$ most recent locations. The $k$ most recent locations correspond to the current state in an underlying stationary order-$k$ Markov source with a transition probability matrix $M$.

Consider an individual whose location history is $L = \{X_1 = a_1, X_2 = a_2 \ldots X_n = a_n\}$, where each $X_i$ is a random variable representing the individual's $i^{th}$ location and each $a_i$ belongs to a set $\mathcal{A}$ comprised of all possible locations. Let substring $X(i,j) = \{X_1, X_{i+1}, \ldots X_j\}$ for any $1 \leq i \leq j \leq n$. The current state for order-$k$ Markov model is then $c_n = \{a_{n-k+1} a_{n-k+2 \ldots} a_n\}$. The probability of the next location being $a$ given location history $L$ is:

$$P(X_{n+1} = a|L) = P(X_{n+1} = a|X(n-k+1, n) = c_n), \quad [5.4]$$

where the notation $P(X_i = a_i| \ldots)$ denotes the probability that $X_i$ takes the value of $a_i$.

This probability corresponds to an entry in the transition probability matrix $M$ associated with the order- $k$ Markov source. That is

$$P(X_{n+1} = a|X(n-k+1, n) = c_n) = M(c_n, c_{n+1}), \quad [5.5]$$

where $c_{n+1} = \{a_{n-k+2 \ldots} a_n a\}$ which is the next state in the Markov chain.

However, since $M$ is not known, it needs to be estimated. The entry of $M$—$M(c_n, c_{n+1})$— can be estimated based on location history $L$ and current prediction context $c_n$ as

$$\hat{P}(X_{n+1} = a | X(n-k+1, n) = c_n) = \widehat{M}(c_n, c_{n+1}) = \frac{N(c_n a, L)}{N(c_n, L)}, \quad [5.6]$$

where $N(t, s)$ is the number of times substring $t$ occurs in string $s$.

The predictor then chooses the location having the highest estimated probability as the next location.

Order-$k$ Markov predictor has found its application in a diverse set of wireless networks for location prediction. With wireless traces left in the Wi-Fi network on the campus of University of North Carolina, Chinchilla et al. (208) modelled the next location choice with both order-1 and order-2 Markov predictor and showed that they could predict with a high accuracy of more than 80%. In (173), multiple location predictors were evaluated with data collected on the Wi-Fi network at Dartmouth College. In the prediction of next location of more than 6,000 users, an order-2 Markov predictor achieved a prediction accuracy level of 63%. Nicholson and Noble (209) applied an order-2 Markov predictor to GPS traces and reported an accuracy level over 70% in next location prediction. An accuracy range from 70% to 95% for the next location was reported by Gambs and Killijian (164) with an order-2 Markov predictor applied to three different data sets comprised of GPS traces. Most recently, with the traces of 500,000 individuals mobile users, Lu et al. (65) predicted the next location with an average accuracy of 91% with an order-1 Markov predictor.

In a number of studies focusing on the evaluation of alternative predictors, the order-$k$ Markov predictor was marked by its high prediction accuracy, easy implementation and low

computational cost (173, 210, 211). Song et al. (173) found that an order-2 Markov predictor performed as well or better than other more complex and more space-consuming compression-based predictors. Petzold et al. (211) compared an order-2 Markov predictor with other next location predictors, namely, Bayesian network, multi-layer perception, Elman net and state predictor from multiple dimensions, including prediction accuracy, stability and computing cost. In an experiment with real mobility data, the Markov predictor demonstrated comparable prediction accuracy (about 80%) to other complex algorithms and, at the same time, was characterized with fast learning and low computational cost. Sigg et al. (210) proposed an alignment approach for location prediction and evaluated it on an individual's GPS traces with comparison to an order-2 Markov predictor, a principal component analysis-based and an independent component analysis-based predictor. The Markov predictor outperformed all other predictors in producing the minimal mean square error in location prediction.

One limitation of order-$k$ Markov predictor is its incapability of predicting any new location that has not been observed before. Yet, studies have shown that the probability of individuals exploring new locations decreases over time (212). Therefore, given some time, all available location alternatives should have been observed. The performance of order-$k$ Markov predictor is expected to improve as observation time becomes longer.

### 5.2.4 Location History and Prediction Accuracy

Studies have consistently shown that prediction accuracy of location history-based predictors is sensitive to the length of location history available. Song et al. (173) found that the prediction accuracy of order-1 and order-2 Markov predictors increased with the length of location history. Katsaros and Manolopoulos (207) showed that, for a set of predictors from the Markov family,

prediction accuracy could be doubled when the length of input location history increased from 100 to 500 (not necessarily unique) locations. Michaelis and Wietfeld (170) observed that, when the input location history length increased from 1 to 8, prediction accuracy climbed from a little over 50% to as high as 95%. In an investigation of the correlation between prediction accuracy and length of location history, Lu et al. (65) showed there was a steady increase in prediction accuracy for Markov predictors. No attempts have been made to uncover the correlation between length of location history and prediction accuracy for subpopulations who may differ in their travel characteristics. It is possible that some user traces are simply less predictable than others and some intrinsic characteristics of a trace may determine its predictability, such as entropy (173).

## 5.3 DATA

### 5.3.1 Data Overview

Please refer to Section 3.1 in this dissertation for a detailed description of the data.

### 5.3.2 Data Prepossessing

Please refer to Section 3.2 in this dissertation for a detailed description of the procedure.

### 5.3.3 Location Representation

A $220 \times 180$, two-dimensional grid of square cells is imposed on the study area, each cell 500 meters on a side. This particular discretization of space is a heuristic choice, and a different size of cells can be adopted. Trip length is considered to be an important factor in determining the cell size. The resultant spatial resolution should be high enough to distinguish the origin cell and

the destination cell of a trip. It's well recognized that people live in denser area make shorter trips (213). Therefore, we refer to location prediction studies conducted at populated U.S. cities (101, 214). Their discretization resulted in cell size ranging from 500 m*500 m to 1000m *1000 m. A cell size of 500 m×500 m is selected in this dissertation. Each of the $N = 39{,}600$ cells is given an index $i = 1,2,3,\ldots,N$. After generating grids, locations represented by latitude and longitude coordinates in the location history are replaced by their corresponding cell ids.

### 5.3.4 Sample Selection

In determining prediction accuracy, predicted locations are compared to those observed ones. Sampled individuals should have sufficiently fine temporal resolution of sightings/location updates. If sightings are too sparse and no location update available for a long time period, the correctness of the predicted location can't be properly evaluated. Therefore, our final sample is selected as those who have at least 10 sightings on the days they were observed, that is, 3,568 individuals in total. Table 5.1 shows comparison between our final sample and the original sample in terms of some statistics characterizing their phone activities. From the table, it is evident that individuals in our final sample generally generated more sightings and correspondingly had shorter interval between consecutive sightings. In addition, selected individuals also were observed for a longer time periods than a typical user in the original sample. The implication of this selection will be discussed in more detail in the last section.

Table 5.1 Comparison between Original Sample and Final Sample

| | Mean of average daily number of sightings | Mean of average time interval (min) | Mean of no. of days observed | Mean of total number of sightings |
|---|---|---|---|---|
| Original | 32.90 | 18.62 | 20.47 | 1606 |
| Final | 168.90 | 14.04 | 52.88 | 8909 |

## 5.4 RESULTS

### 5.4.1 Sample Mobility Overview

Fig. 5.1 shows the cumulative distribution of the total number of activity locations visited by the sample during the study period. The majority of our sample only visited a few dozens of unique locations and only a few cases were observed to visit more than 60 locations. This result is highly consistent with the findings in previous studies: individuals tend to return to a few locations they frequently visited and in time, the probability of visiting a previously unobserved location will decrease (56).

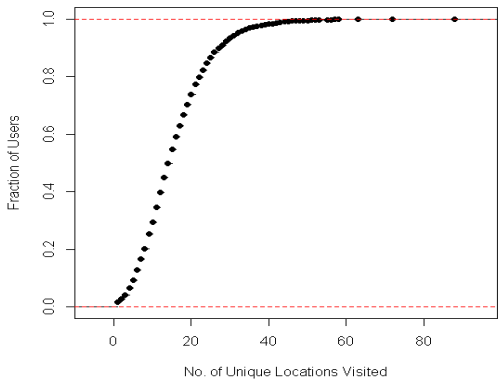


FIGURE 5.1 Cumulative distribution of no. of unique location visited

Moreover, the cumulative distribution of history length is examined (Fig. 5.2). In this dissertation, length of location history describes location change. Location history is defined in a way that it is a sequence of cells an individual has subsequently traversed and thus the next location is not the same as the previous one. Consider a sequence of sights observed for an individual as $\{location\ A_{t_1}, location\ A_{t_2}, location\ B_{t_3}, location\ A_{t_4}, location\ B_{t_5}, location\ A_{t_6}, location\ A_{t_7}$ ...$\}(t_1<t_2<\cdots<t_7)$. The location history of this individual would be $\{location\ A, location\ B, location\ A, location\ B, location\ A$ ...$\}$. Over half of our sample have a location history of fewer than 100 locations, with approximately another half with a history length somewhere between 200 and 300. Only a few outliers can be spotted with over 300 location changes.

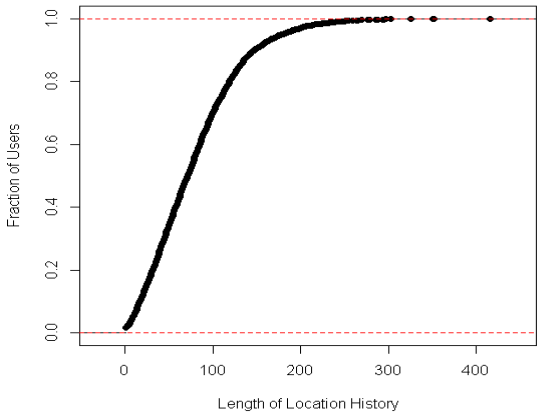


FIGURE 5.2 Cumulative distribution of length of location history

### 5.4.2 Uncertainty and Groups

Fig. 5.3 shows the entropy distribution of our sample. Note that the entropy measured here corresponds to the temporal-uncorrelated entropy measured in (56). Entropy has a bell-shaped distribution with a mean as 3.86 and a range from 0 to a little over 8. These numbers are highly

consistent with those in (56). In order to achieve a balanced group size for comparison purpose, our sample is categorized into four groups based on the quartiles of entropy value. Entropy ranges for these four groups are $[0,3.14)$, $[3.14,3.87)$, $[3.87,4.61)$, $[4.61,8.59]$, respectively.

FIGURE 5.3 Entropy distribution

### 5.4.3 Correlation between History Length and Accuracy

The location predictor used in this dissertation is an order-1 Markov predictor and prediction accuracy is measured as the ratio of the correctly predicted locations to the total predictions made. Fig. 5.4 shows the changes in average prediction accuracy with respect to the length of location history for each group. In general, it is evident that prediction accuracy increases with a longer location history, which is consistent with previous findings (173, 207). Moreover, the most significant improvement in prediction accuracy occurs when the length of location history increases from 10 to 20 historical locations. After that, growth in prediction accuracy gradually slows down and sometimes shows temporary fallback.

The difference between groups is also evident. Given the same history length, groups with low entropy can be predicted more accurately. With only 20 historical locations, we can already achieve an accuracy level over 60% for the group with lowest entropy. When the length of location history reaches 50, accuracy level climbs to over 70%. It keeps increasing to 80% as

history length reaches 100 locations. In contrast, for those individuals who have the highest entropy, prediction accuracy starts at around 40% and slowly increases to approximately 50% after 100 historical locations are observed. For those two groups with modest entropy values, prediction accuracy falls somewhere between 0.6 and 0.7 with an increasing trend towards longer location history.

There are also some unexpected behaviors of the line describing the group with the lowest entropy (the first group): when the length of location history is comprised of less than 40 historical locations, prediction accuracy for this group doesn't exhibit an improvement from that for the group with the second lowest entropy (the second group). A possible explanation is provided here. The sample is grouped based on entropy. Entropy captures the difference in the level of uncertainty/regularity in people's movements based on two-month data. However, this difference in the level of uncertainty/regularity can be very hard to discern for the Markov predicator given short location histories. Markov predictor identifies regular movements by tracking their relative frequency in location history. With limited location history, identification of regular movements would be associated with high uncertainty and lead to inaccurate predictions. But, as location history gets longer, more recurring patterns in the travel of those people can be detected and the difference between the two groups manifests. As shown in Fig. 5.4, a demonstrated improvement in prediction accuracy for the first group over that for the second group is evident after a history length of 50. Especially when the length of location history reaches 100, average prediction accuracy for the first group has reached approximately 0.8, while prediction accuracy for the second group remains at around 0.7.

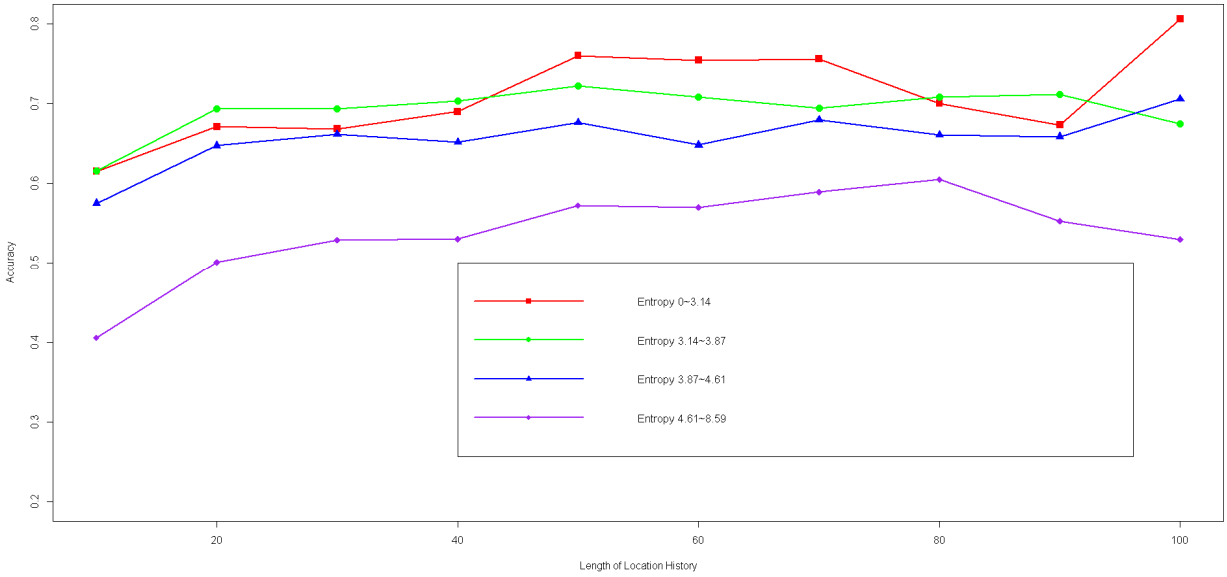


FIGURE 5.4 Correlation between accuracy and history length by groups

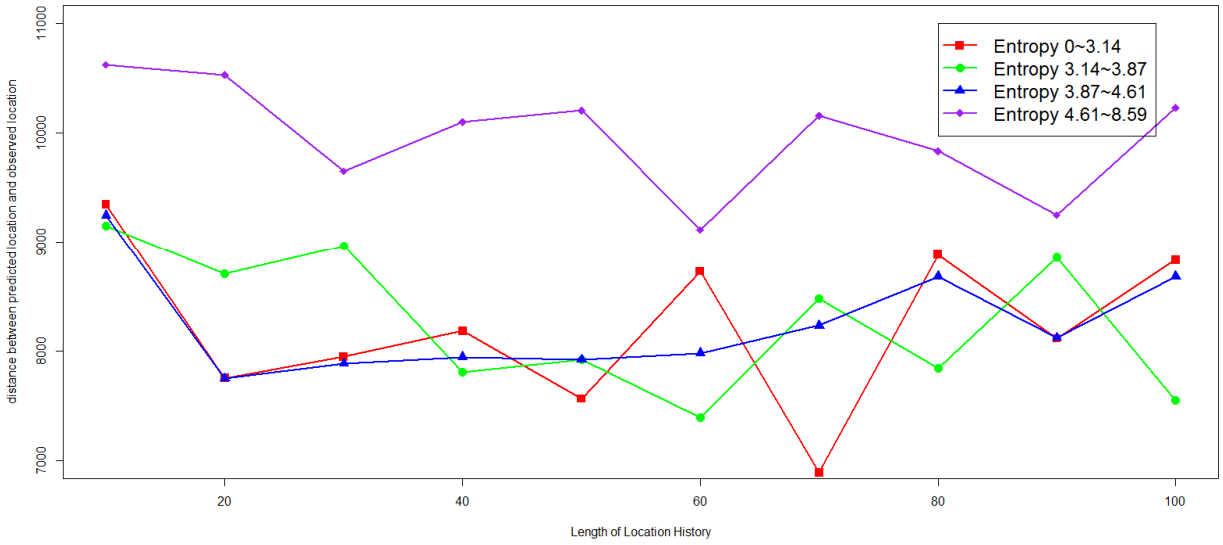


FIGURE 5.5 Correlation between offset distance and history length by groups

In order to further examine the performance of the predictor, the distance between the predicted location and the observed location—the offset distance—is measured and plotted in Fig. 5.5. When a predicted location doesn't match the observed location, a small offset distance is desirable. If an offset distance is smaller enough, the predicted location can still serve as a reasonable representation of the observed location. The general trend in Fig. 5.5 resonates with what is observed in Fig. 5.4. In general, with longer location history, offset distance becomes smaller. The most noticeable difference between groups is those offset distances observed for the group having the highest level of uncertainty and those for the rest of the sample. Those offset distances for the highest-entropy group are all greater 9,000 meters, while the rest of the sample has offset distances generally under 80,000 meters. It seems that, for those people having a high level of uncertainty in their movements, it is difficult not only to pinpoint their locations but also to obtain a close enough predicted location.

## 5.5 CONCLUSION AND DISCUSSIONS

In this chapter, a data set consisting of the traces of 3,568 mobile phone users is used to explore the possibility of increasing the efficiency of location prediction. This issue is approached from two aspects: supporting predictors requiring less computational resource and reducing redundant input information in predictors. It is demonstrated that 1) Markov predictor, characterized with its simple model structure and low computational consumption, is a useful tool in location prediction; 2) the movements of individuals who have low uncertainty in their trajectories require limited amount of input information for a satisfactory prediction.

This dissertation serves as one of a few existing applications of Markov predictor in cellular network. A myriad of location prediction algorithms have been proposed in the past a

few years (20, 215, 216). Selection of predictor always requires a balance between accuracy and efficiency. Comparison studies would aid informed decisions by providing detailed information on computational cost and prediction accuracy for alternative predictors. Song et al. (173) were among the first to evaluate a set of predictors with a real data set. They showed that Markov predictor, despite its simple model structure, outperformed many more complex models in location prediction accuracy. It is worth to note that the data set from Song at al. (173) was from a Wi-Fi network within a limited geographical area (e.g. campus), as from most of the existing studies employing Markov predictors. The power of Markov predictor remains to be tested with data from cellular network covering a large area.

One recent experiment of Markov predictor in cellular network can be found in Lu et al. (65). In their study, the authors described, for different order Markov predictors, the variations in accuracy levels with respect to the length of location histories. One noticeable difference of this study from ours is the definition of location. Location in Lu et al. (65) is defined as a cell in cellular network, without differentiating an activity location, where an individual can spend a significant amount of time, from a transient location (i.e. a cell traversed when an individual travels to the next activity location). On the other hand, of interest in this dissertation is activity location. The distinction between activity location and transient location matters in many cases. Consider the case that location prediction is used to provide weather service. It is probably more meaningful to forecast the weather at the next activity location than a location on the route there. Another difference between these two studies is that Lu et al. (65) didn't report the correlation between prediction accuracy levels and the length of location histories by subgroups differing in the amount of uncertainty in their movements, which is a major contribution of this dissertation.

Entropy, as a measurement of the amount of uncertainty, has been used to quantify the degree of predictability of one's whereabouts in space. Predictability sets the theoretical limit of correct predictions can be made by a location prediction model. Studies have identified rather low entropy values in human trajectories and thus a high level of predictability (56). However, these studies haven't provided a practical location prediction model (130). On the other hand, a myriad of location prediction models have been developed driven by various applications. However, efforts of quantifying predictability and developing location predication model have been largely pursued separately, though predictability serves as the scientific ground for the development of practical predictive models (56).

There has been limited effort to connect entropy with the accuracy level of a practical location prediction model. In Zhao et al. (62), the authors computed entropy values of individual trajectories and divided the population into subgroups based on their entropy values. A best location prediction model was subsequently determined for each subgroup. While Zhao et al. (62) used entropy measure as an indicator for model selection, here, the entropy measure is applied as an indicator of the amount of information input in the predictor.

Our results have important implications on the development of more efficient location predictors. First, it suggests that it is possible to customize the prediction by including varying lengths of location history for different subgroups in the population. It is possible to improve prediction efficiency for those individuals whose movements are fairly regular, as a large portion of the information contained in their location history is redundant and thus dispensable. Being able to discard a significant amount of input information would greatly benefit a wide range of applications relying on accurate and, more importantly, efficient location prediction. Secondly, the Markov predictor is demonstrated to be a powerful tool in making location prediction in

cellular environment. Markov predictor stands out among other location prediction models due to its minimum requirements of memory and computational resource. These two advantages become prominent for location predictions performed in a distributed manner on individual mobile devices with limited memory. In summary, it is possible to develop more efficient location prediction models without compromising prediction accuracy.

This study is not without limitations. One limitation is that the length of the observation period of the data set. For many subjects of our sample, only fewer than a hundred historical locations were recorded, which has limited our investigation on the long-term relationship between history length and prediction accuracy. Yet, prediction accuracy tends to level off after 50 historical locations. There is no obvious reason to expect any significant rise in the accuracy level even a longer location history is not currently available.

Another limitation relates to our sample selection. As discussed above, our findings are derived for a sample who is more engaged in phone activities. This raises the question of representativeness. It is common for researchers to select a non-random study sample from all the subscribers included in a raw mobile phone data set provided by the network operator. As an example, in (56), a sample of mobile phone users who made at least one call every two hours was selected. Recent studies (103, 114, 115) show that user mobility had a strong correlation with phone usage: more active users are more mobile. Therefore, sample selection based on phone usage would potentially result in an overestimation of mobility levels. However, some studies (114) also suggested that some mobility measures seemed to be immune to this sampling bias, such as radius of gyration. While this issue requires further studies, our primary objective here is to compare the prediction accuracy achieved for groups of people differing in the

uncertainty level in their trajectories, rather than to derive an accurate measure to characterize their mobility.

# CONCLUSIONS AND IMPLICATIONS

Data collection and modeling practices in travel behavior research rely heavily on the assumption of repetitious travel behaviors (11). Yet, our own reflection would suggest that we are not repeating the same activity-travel patterns everyday. Though the interest in understanding behavioral variability is not new, research efforts have been greatly hampered by the lack of longitudinal data (106). There has been a renewed interest on this issue due to the availability of new data collection technologies, such as Global Positioning System (GPS)-based travel survey and mobile phone-based data. Originally collected for billing and network maintenance purposes, mobile phone data contains location information of a large portion of population over an extended period of time. It provides travel behavior researchers with an unprecedented opportunity of studying behavior variability.

Previous studies on behavioral variability have concentrated on day-to-day variations in travel behavior, that is, behaviors under comparison are aggregated on a daily basis, such as daily trip rates or daily travel time. This practice precludes insights into behavioral variability at a finer temporal resolution. Consider a person who makes one additional trip on a second day. This observation could have resulted from various activity-travel decisions: it is possible that this person makes an additional trip back home for lunch at noon; it is also likely that this person makes an additional stop at a grocery store before heading home in the evening. The former case leads to an increased travel demand during lunch time, while an increase in travel demand occurs during the evening in the latter case. However, it is impossible to differentiate these two cases if behavioral variability is examined on a daily basis. To discern behavioral variability with temporal resolution finer than one day would require an elaboration on the time of a day when

behavioral variability occurs. There comes my first research question—the time-of-day dependence of location variability.

Research on the time-of-day dependence of location variability is driven by both analytical and policy needs. First, time-of-day plays a critical role in shaping one's location choice behavior in that travel behavior is constrained by not only how fast one can travel but also by the amount of time available for activity and travel (15). Explicitly taking account for the time-of-day effect in modeling location variability would enhance the explanatory power of these models. Secondly, many transportation policies rely on an understanding the influence of time-of-day on location variability, such as time-of-day pricing on major corridors. Less variable location choices in the morning would probably suggest that a higher price is required to alter individuals' location choices during the morning.

Analysis results in this dissertation not only confirm the existence of time-of-day dependence of location variability, but also identify time-of-day as a variable explaining a surprisingly large portion of variations in location variability. Individuals are found to be more likely to vary their location choices in the afternoon periods than in the morning and evening and time-of-day takes account for approximately 36% of the total variations in location variability. These findings all suggest time-of-day is an important factor in influencing individuals' location choices and provide valuable insights into the magnitude of time-of-day effect.

One of the major motivations for understanding activity location choice is to predict individuals' location choices. It is desirable to apply our knowledge of location variability to facilitate location prediction in order to benefit a wide range of practical applications. Therefore, my second research question concerns the connection between location variability and prediction

efficiency. Efficiency of current location history-based predictors is limited by the computational resources used to process large amount of individual location history as model input.

The needs of continuously improving the efficiency of these location predictors primarily come from two aspects. First, location prediction can be performed in two manners: through a central infrastructure or through individual mobile devices. These days, of great concern is the possibility of individuals' location information being compromised if the central infrastructure performing location prediction experiences hardware failure or hacker infiltration (71). There is an increasing need of implementing location prediction in a distributed manner on individual mobile devices. Yet, the limited memory of individual mobile device also poses challenges to the efficiency of location prediction. Second, a wide range of location-based services value more efficient location predictions which allow them to operate in a proactive manner. Imagine a weather service would serve customers better if it can predict your next location well in advance and save you the travel if the weather at your planned activity location is not ideal.

In order to improve the prediction efficiency of location history-based predictors, I ask how much of the input location history is necessary; alternatively, to what extent can this information be compressed. This question stems from two observations. First, a significant amount of regularity in human travel has been identified in recent studies (57). If individuals' travel is fairly regular, location choices observed during a short time period should be sufficient to typify their travel decisions in the long run. Second, for the past decades or so, location prediction in the transportation field has been largely performed with cross-sectional data recording individuals' location information for just one day. Similarly, there is a possibility, with location history-based predictors, that a short location history would suffice for a sufficiently accurate location prediction, at least for those who are less likely to vary their location choices.

A critical finding from this part of my dissertation is that location variability can serve as an instrumental indicator of the amount of input information in location prediction. Specifically, given 100 historical locations, an accuracy level marginally over 80% can be achieved for people with low location variability. In contrast, for those individuals with high level of location variability, prediction accuracy level can hardly reach 50% with 100 historical locations.

My research results have important implications on location prediction practices. First, research results can be used to customize the amount of information input in location prediction. As is shown the Fig. 5.4, for individuals characterized with different levels of uncertainty in their movements, the length of location history required to achieve certain prediction accuracy level differs. For instance, if a prediction accuracy level over 60% is desired, 10 historical locations would suffice for those individuals with lowest level of uncertainty. In other words, it is possible to estimate the exact time point to cut off a location history based on a desired accuracy level. This finding allows us to discard redundant information input for those having little uncertainty in their trajectory, which leads to more efficient computation. Second, Fig. 5.4 shows prediction accuracy levels off after 50 historical locations. For those individuals with the highest level of uncertainty in their trajectories, prediction accuracy stabilizes between 50% and 60%. It appears the high level of uncertainty contained in these individuals' travel would prevent us from making a significant improvement on prediction accuracy level even with a longer location history. If a high level of uncertainty tends to fail any attempt of making highly precise predictions, applications that rely on high location prediction accuracy are expected to be more productive by focusing on those having less uncertainty in movements. In this case, level of uncertainty may serve as an instrumental indicator in helping us to determine the target population for certain applications.

This dissertation induces some interesting work to be explored in future research. First, location variability can be quantified by many measures. For instance, previous studies measured location variability as the percentage of non-repetitive activity locations out of the total number of activity location visited (106). In this dissertation, location variability is measured with entropy. Different quantities capture different dimensions of variability. Schlich & Axhausen (217) compared three measures used to quantify the variability in daily activity-travel patterns and confirmed that the variability increases if the measurement captures more of the complexity of the travel pattern. There is probably no single best measure of location variability and the selection may significantly depend on the kind of applications. Given the limited number of studies accumulated on the variability of spatial behavior, future efforts are needed to introduce other measures of location variability in order to comprehensively characterize one's location variability and generate more insights into one's location choice behavior.

While time-of-day is found to explain a significant amount of variation in location variability, a large portion of variation remains unaccounted for. Further efforts to explain the remaining variation are expected to facilitate the modeling of location variability and deepen our understanding of the mechanisms underlying location variability. A class of variables of potential explanatory power is socio-demographic variables. Subpopulations with varying socio-demographic characteristics have been shown to exhibit different amount of variability in spatial behavior, such as action space (218). Yet, previous results on the relationship between socio-demographic variables and behavioral variability are mixed and the number of socio-demographic variables examined in the context of behavioral variability is limited. As pointed by Kitamura et al. (2006, pp. 269), "This scarcity of explanatory variables is presumably because it has not been customary in the travel behavior analysis field to measure variables that may be

associated with day-to-day variability in travel…". In order to bridge this gap in travel behavior research, a meaningful next-step of this dissertation is to examine the impacts of socio-demographic variables on location variability.

# LIST OF REFERENCES


1.T. Hägerstrand, (1970). What about People in Regional Science? Regional Science, 24, 1, 7-24.
2.J. O. Huff and S. Hanson, (1988). Repetition and Variability in Urban Travel. Geographical Analysis, 18, 2, 97-114.
3.K. G. Goulias, (1999). Longitudinal analysis of activity and travel pattern dynamics using generalized mixed Markov latent class models. Transportation Research Part B, 33, 535-557.
4.M. J. Roorda and T. Ruiz, (2008). Long- and short-term dynamics in activity scheduling: A structural equations approach. Transportation Research Part A, 42, 545–562.
5.A. Simma and K. Axhausen, (2001). Successive days, related travel behaviour ? Arbeitsbericht Verkehrs- und Raumplanung, 62.
6.E. I. Pas, (1987). INTRAPERSONAL VARIABILITY AND MODEL GOODNESS-OF-FIT. Transportation Research Part A, 21A, 6, 431-438.
7.P. L. Mokhtarian and I. Salomon, (2001). How Derived is the Demand for Travel? Some Conceptual and Measurement Considerations. Transportation Research Part A, 35, 695–719.
8.R. Kitamura, T. Yamamoto, Y. O. Susilo and K. W. Axhausen, (2006). How routine is a routine? An analysis of the day-to-day variability in prism vertex location. Transportation Research Part A, 40, 259–279.
9.E. I. Pas, (1988). Weekly travel-activity behavior. Transportation, 15, 89-109.
10.E. I. Pas and S. Sundar, (1995). Intrapersonal variability in daily urban travel behavior:Some additional evidence. Transportation, 22, 135-150.
11.S. Hanson and J. Huff, (1982). Assessing Day-to-Day Variability in Complex Travel Patterns. Transportation Research Record: Journal of the Transportation Research Board, 891, 18-24.
12.E. I. Pas and F. S. Koppelman, (1987). An examination of the determinants of day-to-day variability in individuals' urban travel behavior. Transportation, 14, 3-20.
13.C. R. Bhat and R. Misra, (1999). Discretionary activity time allocation of individuals between in-home and out-of-home and between weekdays and weekends. Transportation, 26, 193-209.
14.J. L. Bowman and M. E. Ben-Akiva, (2000). Activity-based disaggregate travel demand model system with activity schedules. Transportation Research Part A, 35, 1-28.
15.R. Kitamura, C. Chen and R. Narayanan, (1998). Traveler Destination Choice Behavior: Effects of Time of Day, Activity Duration, and Home Location. Transportation Research Record: Journal of the Transportation Research Board, 1645, 1, 76-81.
16.I. G. Cullen and V. Godson, (1975). Urban networks: The structure of activity patterns. 96.
17.P. Jones and M. Clarke, (1988). The significance and measurement of variability in travel behaviour. Transportation, 15, 65-87.
18.P.-R. Lei, T.-J. Shen, W.-C. Peng and I.-J. Su, (2011). Exploring Spatial-Temporal Trajectory Model for Location Prediction. 2011 12th International Conference on Mobile Data Management, IEEE.
19.M. A. Bayir, M. Demirbas and N. Eagle, (2010). Mobility profiler: A framework for discovering mobility profiles of cell phone users. Pervasive and Mobile Computing, 6, 435-454.
20.M. Morzy, (2007). Mining Frequent Trajectories of Moving Objects for Location Prediction. Machine Learning and Data Mining in Pattern Recognition, 4571, 667-680.
21.G. Yavaş, D. Katsaros, Ö. Ulusoy and Y. Manolopoulos, (2005). A data mining approach for location prediction in mobile environments. Data & Knowledge Engineering, 54, 121-146.
22.T. T. Ahonen and A. Moore, (2006). A mobile phone for every living person in Western Europe: penetration hits 100%. communities dominate brands, July 13.

23.D. Wang, D. Pedreschi, C. Song, F. Giannotti and A.-L. Barabási, (2011). Human Mobility, Social Ties, and Link Prediction. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, , 1100-1108.
24.N. Eagle, A. S. Pentland and D. Lazer, (2009). Inferring friendship network structure by using mobile phone data. Proceedings of the National Academy of Sciences, 106, 36, 15274-15278.
25.S. Phithakkitnukoon and R. Dantu, (2011). Mobile social group sizes and scaling ratio. AI & Soc, 26, 71–85.
26.J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész and A.-L. Barabási, (2007). Structure and tie strengths in mobile communication networks. Proceedings of the National Academy of Sciences, 104, 18, 7332-7336.
27.J.-P. Onnela, S. Arbesman, M. C. González, A.-L. Barabási and N. A. Christakis, (2011). Geographic Constraints on Social Network Groups. PloS one, 6, 4, 1-7.
28.F. Calabrese, Z. Smoreda, V. D. Blondel and C. Ratti, (2011). Interplay between Telecommunications and Face-to-Face Interactions: A Study Using Mobile Phone Data. PloS one, 6, 7, 1-6.
29.S. Phithakkitnukoon, F. Calabrese, Z. Smoreda and C. Ratti, (2011). Out of Sight Out of Mind – How our mobile social network changes during migration. IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing, 515-520.
30.S. Phithakkitnukoon, Z. Smoreda and P. Olivier, (2012). Socio-Geography of Human Mobility: A Study Using Longitudinal Mobile Phone Data. PloS one, 7, 6, 1-9.
31.R. A. Becker, R. Cáceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky and C. Volinsky, (2011). A Tale of One City: Using Cellular Network Data for Urban Planning. Pervasive Computing, IEEE 10, 4, 18-26.
32.R. Ahas, A. Aasa, S. Silm and M. Tiru, (2010). Daily rhythms of suburban commuters' movements in the Tallinn metropolitan area: Case study with mobile positioning data. Transportation Research Part C, 18, 45–54.
33.A. P. Wesolowski and N. Eagle, Parameterizing the Dynamics of Slums. AAAI Spring Symposium: Artificial Intelligence for Development, 2010.
34.F. Calabrese, F. C. Pereira, G. D. Lorenzo, L. Liu and C. Ratti, The geography of taste: analyzing cell-phone mobility and social events. Proceedings of IEEE International Conference on Pervasive Computing, 2010.
35.S. Silm and R. Ahas, (2010). The seasonal variability of population in Estonian municipalities. Environment and Planning A, 42, 2527-2546.
36.R. Ahas, A. Aasa, A. Roose, Ü. Mark and S. Silm, (2008). Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. Tourism Management, 29, 469-486.
37.R. Ahas, A. Aasa, Ü. Mark, T. Pae and A. Kull, (2007). Seasonal tourism spaces in Estonia: Case study with mobile positioning data. Tourism Management, 28, 898–910.
38.Y. Asakura and T. Iryo, (2007). Analysis of tourist behaviour based on the tracking data collected using a mobile communication instrument. Transportation Research Part A, 41, 684–690.
39.M. Tiru, E. Saluveer, R. Ahas and A. Aasa, (2010). The Positium Barometer: A Web-Based Tool for Monitoring the Mobility of Tourists. Journal of Urban Technology, 17, 1, 71-89.
40.V. A. Traag, F. Calabrese, A. Browet and F. Morlot, (2011). Social Event Detection in Massive Mobile Phone Data Using Probabilistic Location Inference. IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing, 625-628.
41.D. Quercia, N. Lathia, F. Calabrese, G. D. Lorenzo and J. Crowcroft, (2010). Recommending Social Events from Mobile Phone Location Data. IEEE International Conference on Data Mining, 971-976.
42.H. Bar-Gera, (2007). Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times:A case study from Israel. Transportation Research Part C, 15, 380–391.
43.H. X. Liu, A. Danczyk, R. Brewer and R. Starr, (2008). Evaluation of Cell Phone Traffic Data in Minnesota. Journal of the Transportation Research Board, No. 2086, 1-7.

44.H. Wang, F. Calabrese, G. D. Lorenzo and C. Ratti, (2010). Transportation Mode Inference from Anonymized and Aggregated Mobile Phone Call Detail Records. 13th International IEEE Annual Conference on Intelligent Transportation Systems, 318-323.
45.J. Doyle, P. Hung, D. Kelly, S. Mcloone and R. Farrell, (2011). Utilising Mobile Phone Billing Records for Travel Mode Discovery. ISSC 2011, Trinity College Dublin, June 2011.
46.N. Caceres, J. P. Wideberg and F. G. Benitez, (2007). Deriving origin–destination data from a mobile phone network. IET Intelligent Transport Systems, 1, 1, 15-26.
47.K. Sohn and D. Kim, (2008). Dynamic Origin–Destination Flow Estimation Using Cellular Communication System. IEEE Transactions on Vehicular Technology, 57, 5, 2703-2713.
48.F. Calabrese, G. D. Lorenzo, L. Liu and C. Ratti, (2011). Estimating Origin-Destination Flows Using Mobile Phone Location Data. Pervasive Computing, IEEE, 10, 4, 36–44.
49.J. White and I. Wells, (2002). Extracting Origin Destination Information from Mobile Phone Data. Road TranSport Information and Control, March, 19-21.
50.N. Caceres, L. M. Romero, F. G. Benitez and J. M. D. Castillo, (2012). Traffic Flow Estimation Models Using Cellular Phone Data. IEEE Transactions on Intelligent Transportation Systems, 13, 3, 1430-1441.
51.R. A. Becker, R. Caceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky and C. Volinsky, (2011). Route Classification Using Cellular Handoff Patterns. UbiComp, 123-132.
52.Y. Yim, (2003). The State of Cellular Probes. California PATH Working Paper, UCB-ITS-PRR-2003-25.
53.J. Steenbruggen, M. T. Borzacchiello, P. Nijkamp and H. Scholten, (2011). Mobile phone data from GSM networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities. GeoJournal, DOI 10.1007/s10708-011-9413-y.
54.N. Caceres, J. P. Wideberg and F. G. Benitez, (2008). Review of traffic data estimations extracted from cellular networks. IET Intelligent Transport Systems, 2, 3, 179–192.
55.G. Rose, (2006). Mobile Phones as Traffic Probes: Practices, Prospects and Issues. Transport Reviews, 26, 3, 275-291.
56.C. Song, Z. Qu, N. Blumm and L.-L. Barabási, (2010). Limits of predictability in human mobility. Science, 327, 5968, 1018-1021.
57.M. C. González, C. A. Hidalgo and A.-L. Barabási, (2008). Understanding individual human mobility patterns. Nature, 453, 779-782.
58.F. Calabrese and C. Ratti, (2006). Real time Rome. Networks and Communication studies, 3-4, 247-258.
59.C. Ratti, A. Sevtsuk, S. Huang and R. Pailer, (2005). Mobile Landscapes: Graz in Real Time. Proceedings of the 3rd Symposium on LBS & TeleCartography, 433-444.
60.D. Ettema, H. Timmermans and L. Van Veghel, Effects of Data Collection Methods in Travel and Activity Research. European Institute of Retailing and Services Studies, 1996.
61.C. Song, T. Koren, P. Wang and A.-L. Barabási, (2010). Modelling the scaling properties of human mobility. Nature Physics, 6, 818-823.
62.N. Zhao, W. Huang, G. Song and K. Xie, (2011). Discrete Trajectory Prediction on Mobile Data. Proceedings of the 13th Asia-Pacific web conference on web technologies and applications, 77-88.
63.K. Laasonen, (2005). Route Prediction from Cellular Data. Proceedings of the workshop on Context Awareness for Proactive Systems 147-158.
64.R. Ahas, S. Silm, O. Järv, E. Saluveer and M. Tiru, (2010). Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones. Journal of Urban Technology, 17, 1, 3-27.
65.X. Lu, E. Wetter, N. Bharti, A. J. Tatem and L. Bengtsson, (2013). Approaching the Limit of Predictability in Human Mobility. Scientific Reports, DOI:10.1038/srep02923.
66.X. Lu, L. Bengtsson and P. Holme, (2012). Predictability of population displacement after the 2010 Haiti earthquake. PNAS, 109, 29, 11576-11581.

67.S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, J. Rowland and A. Varshavsky, (2010). A Tale of Two Cities. Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications, ACM, 19-24.
68.S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland and A. Varshavsky, (2011). Ranges of Human Mobility in Los Angeles and New York. Pervasive Computing and Communications Workshops, IEEE, 88-93.
69.K. Laasonen, (2005). Clustering and prediction of mobile user routes from cellular data. Knowledge Discovery in Databases: PKDD, 569-576.
70.S. Akoush and A. Sameh, (2007). Mobile User Movement Prediction Using Bayesian Learning for Neural Networks. Proceedings of the 2007 international conference on Wireless communications and mobile computing, ACM.
71.A. Rodriguez-Carrion, C. Garcia-Rubio, C. Campo and A. Cortés-Martín, (2012). Study of LZ-Based Location Prediction and Its Application to Transportation Recommender Systems. Sensors, 12, 7496-7517.
72.J. Reades, F. Calabrese, A. Sevtsuk and C. Ratti, (2007). Cellular Census: Explorations in Urban Data Collection. Pervasive Computing, IEEE, 6, 3, 30-38.
73.A. Sevtsuk and C. Ratti, (2010). Does Urban mobility have a daily routine? learning from the aggregate data of mobile networks. Journal of Urban Technology, 17, 1, 41-60.
74.R. M. Pulselli, P. Romano, C. Ratti and E. Tiezzi, (2008). Computing Urban Mobile Landscapes Through Monitoring Population Density Based on Cell-Phone Chatting. Int. J. of Design and Nature and Ecodynamics, 3, 2, 121-134.
75.J. B. Sun, J. Yuan, Y. Wang, H. B. Si and X. M. Shan, (2011). Exploring space–time structure of human mobility in urban space. Physica A, 390, 929-942.
76.G. Sagl, M. Loidl and E. Beinat, (2012). A Visual Analytics Approach for Extracting Spatio-Temporal Urban Mobility Information from Mobile Network Traffic. ISPRS International Journal of Geo-Information 1, 3, 256-271.
77.F. Girardin, A. Vaccari, A. Gerber, A. Biderman and C. Ratti, TOWARDS ESTIMATING THE PRESENCE OF VISITORS FROM THE AGGREGATE MOBILE PHONE NETWORK ACTIVITY THEY GENERATE. Intl. Conference on Computers in Urban Planning and Urban Management, 2009.
78.P. A. Pellegrini and A. S. Fotheringham, (2002). Modelling spatial choice: a review and synthesis in a migration context. Progress in Human Geography, 26, 4, 487–510.
79.V. Soto and E. Frías-Martínez, (2011). Automated Land Use Identification using Cell-Phone Records. Proceedings of the 3rd ACM international workshop on MobiArch, 17-22.
80.Y. Yuan and M. Raubal, (2012). Extracting dynamic urban mobility patterns from mobile phone data. Geographic Information Science 354-367.
81.M. R. Vieira, V. Frías-Martínez, N. Oliver and E. Frías-Martínez, (2010). Characterizing Dense Urban Areas from Mobile Phone-Call Data: Discovery and Social Dynamics. Proceedings of IEEE Second International Conference on Social Computing 241-248.
82.W. Huang, Z. Dong, N. Zhao, H. Tian, G. Song, G. Chen, Y. Jiang and K. Xie, (2010). Anchor Points Seeking of Large Urban Crowd Based on the Mobile Billing Data. Advanced Data Mining and Applications, 346-357.
83.B. C. Csáji, A. Browet, V. A. Traag, J.-C. Delvenne, E. Huens, P. V. Dooren, Z. Smoreda and V. D. Blondel, (2013). Exploring the mobility of mobile phone users. Physica A, 392, 1459–1473.
84.S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, Margaretmartonosi, J. Rowland and A. Varshavsky, (2011). Identifying Important Places in People's Lives from Cellular Network Data. Proceedings of International Conference on Pervasive Computing, Jun. 2011.
85.J. Candia, M. C. González, P. Wang, T. Schoenharl, G. Madey and A.-L. Barabási, (2008). Uncovering individual and collective human dynamics from mobile phone records. JOURNAL OF PHYSICS A, 41, 1-11.

86.L. Bengtsson, X. Lu, A. Thorson, R. Garfield and J. V. Schreeb, (2011). Improved Response to Disasters and Outbreaks by Tracking Population Movements with Mobile Phone Network Data: A Post-Earthquake Geospatial Study in Haiti. PloS one, 8, 8, 1-9.
87.J. P. Bagrow, D. Wang and A.-L. Barabási, (2011). Collective Response of Human Populations to Large-Scale Emergencies. PloS one, 6, 3, 1-8.
88.J. P. Bagrow and Y.-R. Lin, (2012). Mesoscopic structure and social aspects of human mobility. PloS one, 7, 5, 1-11.
89.E. Cho, S. A. Myers and J. Leskovec, (2011). Friendship and Mobility: User Movement in Location-Based Social Networks. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, 1082-1090.
90.D. Zhang, D. Zhang, H. Xiong, L. T. Yang and V. Gauthier, (2013). NextCell: Predicting Location Using Social Interplay from Cell Phone Traces. IEEE TRANSACTIONS ON COMPUTERS, DOI 10.1109/TC.2013.223.
91.M. D. Domenico, A. Lima and M. Musolesi, (2012). Interdependence and Predictability of Human Mobility and Social Interactions. Interdependence and Predictability of Human Mobility and Social Interactions, Oct. 2012.
92.J.-L. Ygnace, C. Benguigui and V. Delannoy, Travel Time/Speed Estimates on the French Rhone Corridor Network Using Cellular Phones as Probes. Final Report of the SERTI V Program, 2001.
93.S. Reddy, M. Mun, J. Burke, D. Estrin, M. Hansen and M. Srivastava, (2010). Using Mobile Phones to Determine Transportation Modes. ACM Transactions on Sensor Networks, 6, 2, 13.
94.S. Vlassenroot, R. Bellens, D. Verstraeten and S. Gautama, (2012). The MOVE Project: Smartphones for Smart Travel-behaviour Data Analyses. 19th ITS World Congress, 2012.
95.J. Chen, J. Newman and M. Bierlaire, (2009). Modeling Route Choice Behavior From Smart-phone GPS data. 12th International Conference on Travel Behaviour Research, Jaipur, Rajasthan, India.
96.M. Ficek and L. Kencl, Inter-Call Mobility model: A spatio-temporal refinement of Call Data Records using a Gaussian mixture model. Proceedings of INFOCOM, IEEE, 2012.
97.F. Calabrese, M. Diao, G. D. Lorenzo, J. F. Jr. and C. Ratti, (2013). Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. Transportation Research Part C, 26, 301-313.
98.H. Zang, F. Baccelli and J. Bolot, (2010). Bayesian inference for localization in cellular networks. IEEE INFOCOM 2010 proceedings, 1-9.
99.G. Mao, B. Fidan and B. D. O. Anderson, (2007). Wireless sensor network localization techniques. Computer Networks, 51, 2529–2553.
100.Y. Zhao, (2000). Mobile Phone Location Determination and Its Impact on Intelligent Transportation System. IEEE Transactions, Intelligent Transportation Systems, 1, 55–64.
101.J. Krumm and E. Horvitz, (2006). Predestination: Inferring Destinations from Partial Trajectories. UbiComp 2006: Ubiquitous Computing, 4206, 243-260.
102.Y. Ye, Y. Zheng, Y. Chen, J. Feng and X. Xie, (2009). Mining individual life pattern based on location history. Mobile Data Management: Systems, Services and Middleware, 1-10.
103.C. Iovan, A.-M. Olteanu-Raimond, T. Couronné and Z. Smoreda, (2013). Moving and Calling: Mobile Phone Data Quality Measurements and Spatiotemporal Uncertainty in Human Mobility Studies. Geographic Information Science at the Heart of Europe Lecture Notes in Geoinformation and Cartography, 247-265.
104.J.-K. Lee and J. C. Hou, (2006). Modeling steady-state and transient behaviors of user mobility: formulation, analysis, and application. In Proceedings of the 7th ACM international symposium on Mobile ad hoc networking and computing, ACM, 85-96.
105.D. T. Hartgen and E. S. Jose, Costs and Trip Rates of Recent Household Travel Surveys. November 11, 2009.

106.R. N. Buliung, M. J. Roorda and T. K. Remmel, (2008). Exploring spatial variety in patterns of activity-travel behaviour: initial results from the Toronto Travel-Activity Panel Survey (TTAPS). Transportation Research Part A, 35, 697–722.
107.H. Gong, C. Chen, E. Bialostozky and C. T. Lawson, (2012). A GPS/GIS method for travel mode detection in New York City. Computers, Environment and Urban Systems, 36, 131-139.
108.C. Chen, H. Gong, C. Lawson and E. Bialostozky, (2010). Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the New York City case study. Transportation Research Part A, 44, 830-840.
109.A. J. Richardson, E. S. Ampt and A. H. Meyburg, (1995). Survey Methods for Transport Planning.
110.K. W. Axhausen, M. Löchl, R. Schlich, T. Buhl and P. Widmer, (2007). Fatigue in long-duration travel diaries. Transportation 34, 143–160.
111.S. N. Patel, J. A. Kientz, G. R. Hayes, S. Bhat and G. D. Abowd, (2006). Farther Than You May Think: An Empirical Investigation of the Proximity of Users to Their Mobile Phones. UbiComp 2006: Ubiquitous Computing, 123-140.
112.Xinhua, (2013). Rwanda hits 55pc mobile phone penetration rate. July, http://www.africareview.com/Business---Finance/Rwanda-mobile-phone-penetration-rate/-/979184/1713912/-/format/xhtml/-/dr1a8kz/-/index.html.
113.Experian Simmons, (2011). The 2011 Mobile Consumer Report. July, http://www.experian.com/assets/simmons-research/white-papers/experian-simmons-2011-mobile-consumer-report.pdf.
114.G. Ranjan, H. Zang, Z.-L. Zhang and J. Bolot, (2012). Are Call Detail Records Biased for Sampling Human Mobility? Mobile Computing and Communications Review, 16, 3, 33-44.
115.Z. S. Thomas Couronné, Ana-Maria Olteanu, (2011). Chatty Mobiles: Individual mobility and communication patterns. Analysis of Mobile Phone Datasets and Networks, Oct. 10-11, 2011.
116.D. Schulz, S. Bothe and C. Körner, Human Mobility from GSM Data - A Valid Alternative to GPS? Mobile Data Challenge 2012 Workshop, 2012.
117.Z. Smoreda, A.-M. Olteanu-Raimond and T. Couronné, (2013). Spatiotemporal data from mobile phones for personal mobility assessment. Transport Survey Methods: Best Practice for Decision Making, 1-20.
118.R. Becker, R. Cáceres, K. Hanson, S. Isaacman, J. M. Loh, M. Martonosi, J. Rowland, S. Urbanek, A. Varshavsky and C. Volinsky, (2013). Human Mobility Characterization from Cellular Network Data. Communications of the ACM, 56, 1, 74-82.
119.D. J. Patterson, L. Liao, D. Fox and H. Kautz, (2003). Inferring High-Level Behavior from Low-Level Sensors. UbiComp 2003: Ubiquitous Computing, 73-89.
120.G. D. Lorenzo and F. Calabrese, (2011). Identifying Human Spatio-Temporal Activity Patterns from Mobile-Phone Traces. 14th International IEEE Conference on Intelligent Transportation Systems (ITSC), 1069-1074.
121.S. Phithakkitnukoon, T. Horanont, G. D. Lorenzo, R. Shibasaki and C. Ratti, (2010). Activity-Aware Map: Identifying Human Daily Activity Pattern Using Mobile Phone Data. In: Workshop on Human Behavior Understanding, 14-25.
122.V. Frias-Martinez and J. Virseda, (2012). On The Relationship Between Socio-Economic Factors and Cell Phone Usage. Proceedings of the Fifth International Conference on Information and Communication Technologies and Development. ACM, 76-84.
123.V. Soto, V. Frias-Martinez, J. Virseda and E. Frias-Martinez, (2011). Prediction of Socioeconomic Levels Using Cell Phone Records. User Modeling, Adaption and Personalization, 377-388.
124.K. W. Axhausen, (2002). A dynamic understanding of travel demand: a sketch. Arbeitsberichte Verkehrs- und Raumplanung, 119, Institut für Verkehrsplanung, Transporttechnik, Strassen- und Eisenbahnbau (IVT), ETH Zurich, Zurich.

125.J. A. Carrasco, B. Hogan, B. Wellman and E. J. Miller, (2008). Collecting social network data to study social activity-travel behavior: an egocentric approach. Environment and Planning B, 35, 961-980.
126.C. R. Bhat, S. Srinivasan and K. W. Axhausen, (2005). An analysis of multiple interepisode durations using a unifying multivariate hazard model. Transportation Research Part B, 39, 797–823.
127.C. Chen and J. Chen, (2009). What is responsible for the response lag of a significant change in discretionary time use: the built environment, family and social obligations, temporal constraints, or a psychological delay factor? Transportation, 36, 1, 27-46.
128.O. Järv, R. Ahas and F. Witlox, (2013). Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone call detail records. Transportation Research Part C, 38, 122-135.
129.S. Bekhor, Y. Cohen and C. Solomon, (2013). Evaluating long-distance travel patterns in Israel by tracking cellular phone positions. Journal of Advanced Transportation, 47, 4, 435-446.
130.F. Calabrese, D. L. G. and C. Ratti, (2010). Human mobility prediction based on individual and collective geographical preferences. Proceedings of 13th International IEEE Conference on Intelligent Transportation Systems (ITSC), 312-317.
131.H. Kim and M. Kwan, (2003). Space-time accessibility measures: A geocomputational algorithm with a focus on the feasible opportunity set and possible activity duration. Journal of Geographical Systems, 5, 71-91.
132.A. Santos, N. Mcguckin, H. Y. Nakamoto, D. Gray and S. Liss, (2011). Summary of Travel Trends: 2009 National Household Travel Survey. FHWA-PL-ll-022, Federal Highway Administration.
133.R. Kitamura and T. Van Der Hoorn, (1987). Regularity and irreversibility of weekly travel behavior. Transportation, 14, 227-251.
134.S. Schönfelder, (2001). Some notes on space, location and travel behaviour. 1st Swiss Transport Research Conference.
135.S. Schönfelder and K. W. Axhausen, (2004). Structure and innovation of human activity spaces. Arbeitsberichte Verkehrs-und Raumplanung, 258.
136.Y.-H. Wu and H. J. Miller, (2001). Computational tools for measuring space-time accessibility within dynamic flow transportation networks. Journal of Transportation and Statistics, 4, 2/3, 1-14.
137.M.-P. Kwan, (2000). Gender differences in space-time constraints. Area, 32, 2, 145-156.
138.T. Schwanen and M. Dijst, (2003). Time windows in workers' activity patterns: Empirical evidence from the Netherlands. Transportation, 30, 261–283.
139.A. F. Clark and S. T. Doherty, (2009). Activity Rescheduling Strategies and Decision Processes in Day-to-Day Life. Transportation Research Record: Journal of the Transportation Research Board, 2134, 143-152.
140.S. T. Doherty and E. J. Miller, (2000). A computerized household activity scheduling survey. Transportation, 27, 75–97.
141.H. Kang, D. M. Scott and S. T. Doherty, (2009). Investigation of Planning Priority of Joint Activities in Household Activity-Scheduling Process. Transportation Research Record: Journal of the Transportation Research Board, 2134, 82-88.
142.C. E. Shannon and W. Weaver, (1948). A mathematical theory of communication. The Bell System Technical Journal, 27, 379-423, 623-656.
143.M. Vanhoy, (2008). An Entropy Estimator of Population Variability in Nominal Data. Journal of Scientific Psychology, June, 25-30.
144.C. Fraley and A. E. Raftery, (1998). How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. THE COMPUTER JOURNAL, 41, 8, 578-588.
145.J. D. Banfield and A. E. Raftery, (1993). Model-Based Gaussian and Non-Gaussian Clustering. Biometrics, 49, 3, 803-821.
146.G. Celeux and G. Govaert, (1995). GAUSSIAN PARSIMONIOUS CLUSTERING MODELS. Pattern Recognition, 28, 5, 781-793.

147.A. Dasgupta and A. E. Raftery, (1998). Detecting Features in Spatial Point Processes with Clutter via Model-Based Clustering. Journal of the American Statistical Association, 93, 441, 294-302.
148.K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery and W. L. Ruzzo, (2001). Model-based clustering and data transformations for gene expression data. Bioinformatics, 17, 10, 977-987.
149.R. A. Redner and H. F. Walker, (1984). Mixture densities, maximum likelihood and the EM algorithm. SIAM review, 26, 2, 195-239.
150.C.-H. Joh, S. T. Doherty and J. W. Polak, (2005). Analysis of Factors Affecting the Frequency and Type of Activity Schedule Modification. Transportation Research Record: Journal of the Transportation Research Board, 1926, 19-25.
151.S. Hanson and J. Huff, (1986). Classification issues in the analysis of complex travel behavior. Transportation, 13, 271-293.
152.H. Jeung, Q. Liu, H. T. Shen and X. Zhou, (2008). A Hybrid Prediction Model for Moving Objects. ICDE 2008, IEEE.
153.N. Eagle and A. S. Pentland, (2009). Eigenbehaviors: identifying structure in routine. Behavioral Ecology and Sociobiology, 63, 1057–1066.
154.G. Liu and G. M. Jr., (1996). A class of mobile motion prediction algorithms for wireless mobile computing and communications. Mobile Networks and Applications, 1, 113-121.
155.S. K. Das and S. K. Sen, (1999). Adaptive location prediction strategies based on a hierarchical network model in a cellular mobile environment. The Computer Journal 42, 6, 473-486.
156.A. Bhattacharya and S. K. Das, (2002). LeZi-Update: An Information-Theoretic Framework for Personal Mobility Tracking in PCS Networks. Wireless Networks, 8, 121-135.
157.F. Yu and V. Leung, (2002). Mobility-based predictive call admission control and bandwidth reservation in wireless cellular networks. Computer Networks, 38, 577–589.
158.L. Liao, D. J. Patterson, D. Fox and H. Kautz, (2007). Learning and inferring transportation routines. Artificial Intelligence, 171, 311–331.
159.T. H. N. Vu, K. H. Ryu and N. Park, (2009). A method for predicting future location of mobile user for location-based services system. Computers & Industrial Engineering, 57, 91-105.
160.T. Anagnostopoulos, C. Anagnostopoulos and S. Hadjiefthymiades, (2010). An Online Adaptive Model for Location Prediction. Autonomic Computing and Communications Systems, 64-78.
161.C. Cheng, R. Jain and E. V. D. Berg, (2003). Location Prediction Algorithms for Mobile Wireless Systems. Wireless internet handbook, CRC Press, Inc.
162.H. Abu-Ghazaleh and A. S. Alfa, (2010). Application of Mobility Prediction in Wireless Networks Using Markov Renewal Theory. IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, 59, 2, 788-802.
163.H. Si, Y. W. J. Yuan and X. Shan, (2010). Mobility Prediction in Cellular Network Using Hidden Markov Model. Consumer Communications and Networking Conference (CCNC), IEEE.
164.S. Gambs and M.-O. Killijian, (2012). Next Place Prediction using Mobility Markov Chains. Proceedings of the First Workshop on Measurement, Privacy, and Mobility, ACM.
165.L. Chen, M. Lv and G. Chen, (2010). A system for destination and future route prediction based on trajectory mining. Pervasive and Mobile Computing, 6, 657-676.
166.P. Bilurkar, N. Rao, G. Krishna and R. Jain, (2002). APPLICATION OF NEURAL NETWORK TECHNIQUES FOR LOCATION PREDICATION IN MOBILE NETWORKING. Proceedings of the 9th International Conference on Neural Information Processing, 5, 2157-2161.
167.D. J. Kadhim, T. M. Ali and F. A. Mustafa, (2013). LOCATION PREDICTION IN CELLULAR NETWORK USING NEURAL NETWORK. INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING & TECHNOLOGY, 4, 4, 321-332.
168.K. Majumdar and N. Das, (2003). Neural Networks tor Location Management in Mobile Cellular Communication Networks. Conference on Convergent Technologies for Asia-Pacific Region, IEEE, 2, 647-651.

169.N. Samaan and A. Karmouch, (2005). A Mobility Prediction Architecture Based on Contextual Knowledge and Spatial Conceptual Maps. IEEE TRANSACTIONS ON MOBILE COMPUTING, 4, 6, 537-551.
170.S. Michaelis and C. Wietfeld, (2006). Comparison of User Mobility Pattern Prediction Algorithms to increase Handover Trigger Accuracy. Vehicular Technology Conference, 2, IEEE.
171.D. Katsaros, A. Nanopoulos, M. Karakaya, G. Yavas, Ö. Ulusoy and Y. Manolopoulos, (2003). Clustering Mobile Trajectories for Resource Allocation in Mobile Environments. Advances in Intelligent Data Analysis V, 319-329.
172.M. Morzy, (2006). Prediction of Moving Object Location Based on Frequent Trajectories. Computer and Information Sciences – ISCIS 2006, 4263, 583-592.
173.L. Song, D. Kotz, R. Jain and X. He, (2006). Evaluating Next-Cell Predictors with Extensive Wi-Fi Mobility Data. IEEE TRANSACTIONS ON MOBILE COMPUTING, 5, 12, 1633-1649.
174.M. E. Ben-Akiva and S. R. Lerman, (1985). Discrete Choice Analysis: Theory and Application to Predict Travel Demand. 390.
175.D. Mcfadden, (1978). Modeling the Choice of Residential Location. Institute of Transportation Studies, University of California, 75-96.
176.S. Bekhor and J. N. Prashker, (2008). GEV-based destination choice models that account for unobserved similarities among alternatives. Transportation Research Part B, 42, 243-262.
177.V. L. Bernardin, F. K. Jr. and D. Boyce, (2009). Enhanced Destination Choice Models Incorporating Agglomeration Related to Trip Chaining While Controlling for Spatial Competition. Transportation Research Record: Journal of the Transportation Research Board, 2132, 143-151.
178.K. E. Train, (2003). Discrete Choice Methods with Simulation. 334.
179.A. S. Fotheringham, (1983). Some theoretical aspects of destination choice and their relevance to production-constrained gravity models. Environment and Planning A, 15, 8, 1121-1132.
180.A. S. Fotheringham, (1986). Modellingh ierarchicadle stinationc hoice. Environment and Planning A, 18, 401-418.
181.A. S. Fotheringham, (1988). Note—Consumer Store Choice and Choice Set Definition. Marketing Science, 7, 3, 299-310.
182.A. S. Fotheringham, T. Nakaya, K. Yano, S. Openshaw and Y. Ishikawa, (2001). Hierarchical destination choice and spatial interaction modelling: a simulation experiment. Environment and Planning A, 33, 901 - 920.
183.K. Haynes and A. S. Fotheringham, (1990). The impact of space on the application of discrete choice models. Review of Regional Studies, 2, 39-49.
184.P. A. Pellegrini and A. S. Fotheringham, (1999). Intermetropolitan migration and hierarchical destination choice: a disaggregate analysis from the US Public Use Microdata Samples. Environment and Planning A, 31, 1093-1118.
185.P. A. Pellegrini, A. S. Fotheringham and G. Lin, (1997). An empirical evaluation of parameter sensitivity to choice set definition in shopping destination choice modelws. Regional Science, 76, 2, 257-284.
186.A. Borgers and H. Timmermans, (1987). Choice model specification, substitution and spatial structure effects. Regional Science and Urban Economics, 17, 29-47.
187.L. Lo, (1991). Substitutability, Spatial Structure, and Spatial Interaction. Geographical Analysis, 23, 2, 132-146.
188.R. J. Meyer and T. C. Eagle, (1982). Context-Induced Parameter Instability in a Disaggregate-Stochastic Model of Store Choice. Journal of Marketing Research, 19, 1, 62-71.
189.L. M. Hunt, B. Boots and P. S. Kanaroglou, (2004). Spatial choice modelling: new opportunities to incorporate space into substitution patterns. Progress in Human Geography, 28, 746-766.
190.K. A. Small, (1987). A discrete choice model for ordered alternatives. Econometrica, 55, 2, 409-424.

191.M. Ben-Akiva and M. Bierlaire, (1999). Discrete Choice Methods and their Applications to Short Term Travel Decisions. International Series in Operations Research & Management Science, 23, 5-33.
192.A. Papola, (2004). Some developments on the cross-nested logit model. Transportation Research Part B, 38, 833-851.
193.J. N. Prashker and S. Bekhor, (1998). Investigation of Stochastic Network Loading Procedures. Transportation Research Record: Journal of the Transportation Research Board, 1645, 94-102.
194.P. Vovsha and S. Bekhor, (1998). Link-Nested Logit Model of Route Choice. Transportation Research Record: Journal of the Transportation Research Board, 1645, 133-142.
195.C. R. Bhat, (1998). ANALYSIS OF TRAVEL MODE AND DEPARTURE TIME CHOICE FOR URBAN SHOPPING TRIPS. Transportation Research Part B, 32, 6, 361-371.
196.F. S. Koppelman and C.-H. Wen, (2000). The paired combinatorial logit model: properties, estimation and application. Transportation Research Part B, 34, 75-89.
197.C.-H. Wen and F. S. Koppleman, (2001). The generalized nested logit model. Transportation Research Part B, 35, 627-641.
198.J. Swait, (2001). Choice set generation within the generalized extreme value family of discrete choie models. Transportation Research Part B, 35, 643-666.
199.C. Chen, J. Chen and H. Timmermans, (2009). Historical deposition influence in residential location decisions: a distance-based GEV model for spatial correlation. Environment and Planning A, 41, 2760-2777.
200.C. R. Bhat and J. Guo, (2004). A mixed spatially correlated logit model: formulation and application to residential choice modeling. Transportation Research Part B, 38, 147–168.
201.J. Walker, (2002). Mixed logit (or logit kernel) model: Dispelling misconceptions of identification. Transportation Research Record: Journal of the Transportation Research Board, 1805, 86-98.
202.J. Walker and M. Ben-Akiva, (2002). Generalized random utility model. Mathematical Social Sciences, 43, 303–343.
203.C. R. Bhat, (2003). Random utility-based discrete choice models for travel demand analysis. Transportation Systems Planning: Methods and Applications, 10, 1-30.
204.A. Kemperman, A. Borgers and H. Timmermans, (2002). Incorporating Variety Seeking and Seasonality in Stated Preference Modeling of Leisure Trip Destination. Transportation Research Record: Journal of the Transportation Research Board, 1807, 67-76.
205.A. Sivakumar and C. R. Bhat, (2007). Comprehensive, Unified Framework for Analyzing Spatial Location Choice. Transportation Research Record: Journal of the Transportation Research Board, 2003, 103-111.
206.C. Chen and H. Lin, (2011). Decomposing residential self-selection via a life-course perspective. Environment and Planning Part A, 43, 11, 2608-2625.
207.D. Katsaros and Y. Manolopoulos, PREDICTION IN WIRELESS NETWORKS BY MARKOV CHAINS. IEEE Wireless Communications, April, 2009.
208.F. Chinchilla, M. Lindsey and M. Papadopouli, (2004). Analysis of Wireless Information Locality and Association Patterns in a Campus. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies, 21, IEEE.
209.A. J. Nicholson and B. D. Noble, (2008). BreadCrumbs: Forecasting Mobile Connectivity. Proceedings of the 14th ACM international conference on Mobile computing and networking, ACM.
210.S. Sigg, S. Haseloff and K. David, (2010). An Alignment Approach for Context Prediction Tasks in UbiComp Environments. Pervasive Computing, 9, 4, 90-97.
211.J. Petzold, F. Bagci, W. Trumler and T. Ungerer, (2006). Comparison of Different Methods for Next Location Prediction. Euro-Par 2006 Parallel Processing, 909-918.
212.C. Song, Z. Qu, N. Blumm and A.-L. Barabási, (2010). Limits of Predictability in Human Mobility. Science, 327, 1018, 1018-1021.

213.R. Cervero and K. Kockelman, (1997). TRAVEL DEMAND AND THE 3Ds: DENSITY, DIVERSITY, AND DESIGN. Transportation Research Part D, 2, 3, 199-219.
214.F. Calabrese, G. D. Lorenzo, L. Liu and C. Ratti, (2011). Estimating Origin-Destination Flows Using Mobile Phone Location Data. Pervasive Computing, 10, 4, 36-44.
215.A. Monreale, F. Pinelli, R. Trasarti and F. Giannotti, (2009). WhereNext: a Location Predictor on Trajectory Pattern Mining. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM.
216.J. J.-C. Ying, W.-C. Lee, T.-C. Weng and V. S. Tseng, (2011). Semantic Trajectory Mining for Location Prediction. Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 34-43.
217.R. Schlich and K. W. Axhausen, (2003). Habitual travel behaviour: Evidence from a six-week travel diary. Transportation, 30, 13-36.
218.Y. O. Susilo and R. Kitamura, (2005). Analysis of Day-to-Day Variability in an Individual's Action Space. Transportation Research Record: Journal of the Transportation Research Board, 1902, 124–133.