Understanding GPS and Mobile Phone Data for Origin-Destination Analysis

OCTOBER 2017

U.S. Department of Transportation
Federal Highway Administration

TMIP
Better Methods. Better Outcomes.

Notice

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The U.S. Government assumes no liability for the use of the information contained in this document.

The U.S. Government does not endorse products or manufacturers. Trademarks or manufacturers' names appear in this report only because they are considered essential to the objective of the document.

Quality Assurance Statement

The Federal Highway Administration (FHWA) provides high-quality information to serve Government, industry, and the public in a manner that promotes public understanding. Standards and policies are used to ensure and maximize the quality, objectivity, utility, and integrity of its information. The FHWA periodically reviews quality issues and adjusts its programs and processes to ensure continuous quality improvement.

1. Report No. FHWA-HEP-19-027	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Understanding GPS and Mobile Phone Data for Origin-Destination Analysis		5. Report Date October 31, 2017
		6. Performing Organization Code
7. Authors Cynthia Chen, Xuegang (Jeff) Ban, Feilong Wang, Jingxing Wang, Choudhury Siddique, Rong Fan, Jaehun Lee		8. Performing Organization Report No.
9. Performing Organization Name and Address Department of Civil and Environmental Engineering University of Washington		10. Work Unit No. (TRAIS)
		11. Contract or Grant No. DTFH61-12-D-00013
12. Sponsoring Agency Name and Address United States Department of Transportation Federal Highway Administration 1200 New Jersey Ave. SE Washington, DC 20590		13. Type of Report and Period Covered June 2017 to October 2017
		14. Sponsoring Agency Code HEPP-30

15. Supplementary Notes

The project was managed by Task Manager for Federal Highway Administration, Sarah Sun, who provided detailed technical directions.

16. Abstract

Emerging datasets such as mobile phone and GPS data have now become a promising data source for many transportation planning applications, including origin-destination (OD) analyses, which serve as the basis for transportation investment and policy decisions. Generated from an entirely different process from the traditional household travel surveys, these datasets possess many characteristics that together can affect the accuracy and representativeness of the derived results such as OD matrices greatly. The aim of this report is not to develop methods for OD analysis but to gain a thorough understanding of such emerging datasets. Two datasets are studied in the study: a mobile phone dataset and a GPS dataset. The study results demonstrate the many different characteristics possessed by the two and their implications for OD analysis are discussed.

17. Key Words	18. Distribution Statement
Big Data, Mobile Phone Data, GPS Data, Data Representativeness, Origin Destination Demand Matrices	No restrictions.

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	149	N/A

SI* (MODERN METRIC) CONVERSION FACTORS

APPROXIMATE CONVERSIONS TO SI UNITS

Symbol	When You Know	Multiply By	To Find	Symbol
		LENGTH		
in	inches	25.4	millimeters	mm
ft	feet	0.305	meters	m
yd	yards	0.914	meters	m
mi	miles	1.61	kilometers	km
		AREA		
in^2	square inches	645.2	square millimeters	mm^2
ft^2	square feet	0.093	square meters	m^2
yd^2	square yard	0.836	square meters	m^2
ac	acres	0.405	hectares	ha
mi^2	square miles	2.59	square kilometers	km^2
		VOLUME		
fl oz	fluid ounces	29.57	milliliters	mL
gal	gallons	3.785	liters	L
ft^3	cubic feet	0.028	cubic meters	m^3
yd^3	cubic yards	0.765	cubic meters	m^3
		NOTE: volumes greater than 1000 L shall be shown in m^3		
		MASS		
oz	ounces	28.35	grams	g
lb	pounds	0.454	kilograms	kg
T	short tons (2000 lb)	0.907	megagrams (or "metric ton")	Mg (or "t")
		TEMPERATURE (exact degrees)		
°F	Fahrenheit	5 (F-32)/9 or (F-32)/1.8	Celsius	°C
		ILLUMINATION		
fc	foot-candles	10.76	lux	lx
fl	foot-Lamberts	3.426	candela/m^2	cd/m^2
		FORCE and PRESSURE or STRESS		
lbf	poundforce	4.45	newtons	N
lbf/in^2	poundforce per square inch	6.89	kilopascals	kPa

APPROXIMATE CONVERSIONS FROM SI UNITS

Symbol	When You Know	Multiply By	To Find	Symbol
		LENGTH		
mm	millimeters	0.039	inches	in
m	meters	3.28	feet	ft
m	meters	1.09	yards	yd
km	kilometers	0.621	miles	mi
		AREA		
mm^2	square millimeters	0.0016	square inches	in^2
m^2	square meters	10.764	square feet	ft^2
m^2	square meters	1.195	square yards	yd^2
ha	hectares	2.47	acres	ac
km^2	square kilometers	0.386	square miles	mi^2
		VOLUME		
mL	milliliters	0.034	fluid ounces	fl oz
L	liters	0.264	gallons	gal
m^3	cubic meters	35.314	cubic feet	ft^3
m^3	cubic meters	1.307	cubic yards	yd^3
		MASS		
g	grams	0.035	ounces	oz
kg	kilograms	2.202	pounds	lb
Mg (or "t")	megagrams (or "metric ton")	1.103	short tons (2000 lb)	T
		TEMPERATURE (exact degrees)		
°C	Celsius	1.8C+32	Fahrenheit	°F
		ILLUMINATION		
lx	lux	0.0929	foot-candles	fc
cd/m^2	candela/m^2	0.2919	foot-Lamberts	fl
		FORCE and PRESSURE or STRESS		
N	newtons	0.225	poundforce	lbf
kPa	kilopascals	0.145	poundforce per square inch	lbf/in^2

*SI is the symbol for the International System of Units. Appropriate rounding should be made to comply with Section 4 of ASTM E380.
(Revised March 2003)

Understanding GPS and Mobile Phone Data for Origin-Destination Analysis

October 2017

Prepared for:

Federal Highway Administration

Table of Contents

List of Figures

List of Tables

List of Abbreviations and Symbols

Abbreviations

CDR	call-detail record
CTPP	Census Transportation Planning Package
DBSCAN	Density-based Spatial Clustering
FHWA	Federal Highway Administration
HICOMP	highway congestion monitoring program
HOV	high-occupancy vehicle
MCTC	Moore County Transportation Committee
MPD	mobile phone data
MPO	metropolitan planning organization
PATH	Partners for Advanced Transit and Highways
PSRC	Puget Sound Regional Council
SVM	support vector machines
TAZ	traffic analysis zone
TNMUG	Tennessee Model Users Group
VID	vehicle identification
VUR	vehicle usage ratio

Symbols

φ	temporal resolution φ of a trajectory
R_L	radius of outputting clusters
t	observation period
r_g	radius of gyration of activity
L_c	cluster center
$N_{O,i}$	number of TAZs that have zero trips originated from TAZ (i)
$N_{D,i}$	number of TAZs that have zero trips destined to TAZ (i)
N	total number of TAZ in the study area (360)
$P_{car\ drive\ alone}$	probabilities that travelers in zone i drive alone
$P_{carpool}(i)$	probabilities that travelers in zone i share a car
S	average carpool size
N_k	the number of users in zone k
$T_{ij}(n)$	total number of trips that user n made between zone i and zone j in the observational period
$N_{pop}(i)$	population in zone i
$N_{user}(i)$	number of selected mobile phone users in zone i
V_k	number of users in zone k
W	daily trip production for the entire population
A	number of zones

1.0 Executive Summary

Origin-Destination (OD) estimation for a metropolitan region is a critical component in the regional transportation planning process. OD estimation requires time- and location-stamped data from which individual movement patterns can be inferred. Aside from traditional data sources such as household travel surveys and license plate surveys (see Volume 1), an increasing list of datasets, i.e., the so-called "big data", has emerged due to the prevalence of devices and services that can now provide information on when and where a user is.

In this volume, the emerging datasets are broadly defined as those of passive solicitation, meaning that such data is not generated for transportation applications (or more specifically in the context of this project, for OD estimation), but is a by-product of other non-transportation-related purposes. This volume focuses on two of those datasets: anonymously generated mobile phone data and global positioning system (GPS) data.

1. Anonymously generated mobile phone data: the primary purpose is for billing and operational purposes by mobile phone companies.
2. GPS data: the primary purpose is to provide navigation or other types of location-based services to users.

The purpose of this volume is to investigate the various properties of the two types of datasets listed above, especially those related to the general spatial and temporal patterns of the data, trips, and OD demands. For this, an analysis framework is developed in this volume, including the investigation of three categories of data properties: *zeroth-order*, *first-order*, and *second-order* properties. Zeroth order describes general spatial and temporal aspects of the dataset; first order describes properties related to single activity locations, or trip ends; and second order illustrates properties related to trips (two linked activity locations or trip ends). This volume starts with state of the practice reviews of mobile phone data and GPS data, as well as existing OD estimation methods using the two data sources. Using a mobile phone dataset and a GPS dataset collected from the real world, in-depth investigations and analyses of the various properties of the two datasets are then conducted. This volume concludes with comparisons of the similarities and differences of the properties generated via the two datasets, observations from the comparisons, and their implications to OD analysis using such emerging datasets in transportation.

This volume reveals that specific characteristics of transportation big data may contribute to different data properties when the data is used for OD related analysis. These characteristics may be categorized as the following:

1. *Updating frequency of the random vehicle or device ID*. Raw data are rarely available and each processed data record is often associated with a random ID. How often this random ID is updated has important implications to OD inference. For example, if an ID is updated frequently (e.g., once every day), home and work locations cannot be reliably identified.
2. *Location accuracy and uncertainty.* Low location accuracy can lead to high location uncertainty, which will pose challenges when identifying trip ends and consequently affect OD estimation. The mobile phone data studied in this volume is much coarse in their location estimation than the GPS data, resulting in location estimation with an error range

in the 100s of meters as opposed to within 10 meters for the GPS data. It is also important to note that not all transportation planning applications require a high location accuracy. For example, OD analysis is usually conducted at the zone level, thus the accuracy of mobile phone data may be sufficient for this purpose. On the other hand, high location accuracy, although helpful for trip end identification and related analysis, may also introduce challenges due to other possible considerations such as data privacy/security.

3. *Missing trips.* This is a common problem found for both datasets for very different reasons though. Consequently the types of trips missed are likely different: non-vehicle based trips and longer trips for GPS data and morning trips and shorter trips for the mobile phone data.
4. *Size of the study area.* The size of the study area is crucial: if the area is too small, the data cannot capture the complete trip of a traveler (or vehicle ID), leading to significant underestimation of some of the properties such as trip rates and trip durations. Therefore, for OD-related analysis, data from a relatively large area (such as an entire city or a region) should be used.
5. *Duration of the collected data.* Both datasets show clear weekly patterns and daily variations. As a result, for OD-related analysis, data of at least a few weeks should be analyzed (for mobile phone data, the threshold appears to be at 3 weeks). For trend analysis however, data of longer than 3 weeks are required.
6. *The data generation process.* Although in general it is hard, if not impossible, to know exactly how the data are generated, even some basic investigation and understanding of the data generation process may be helpful. For example, mobile phone data are generated mainly for phone activities, which may or may not be trip related. GPS data, if collected from vehicle navigation devices or monitoring systems, are more likely related to vehicular trips. Such knowledge is relatively easy to obtain, which can nevertheless provide useful insight and explanations to the patterns of the properties discovered from the data.

It is expected that the various characteristics discovered here may also be relevant to other emerging datasets (e.g., app based data) that can be potentially used for planning purposes such as OD estimation. Issues that arise from these characteristics all converge to the fundamental question: does the dataset at hand represent the travel patterns of a region? This *representative* issue should be the central question when analyzing and using emerging big data for OD estimation. The very first task in answering this question is to obtain a thorough understanding of the data at hand. For this, the analysis framework (zeroth, first, and second order properties) proposed in the study can be readily applied to other emerging datasets. In the long term, future research is critically needed to develop tractable methods to address those issues to produce more accurate and reliable trip/OD related information. Below we summarize a list of lessons learned from this study and these lessons may be useful for those who are interested in using emerging datasets in transportation.

- *Data processing*. Data processing is critically needed before activity locations/trip ends can be identified. This step is not about clustering sightings into locations but about removing and correcting noisy records. This is particularly the case for the mobile phone data, which means removing and correcting oscillation sightings (sightings resulted purely from signaling events not indicative of user movements).

- *Activity location/trip identification*. Methods used to identify activity locations or trips ends can be completely different for different datasets as shown for mobile phone and GPS data. The difference is due to the different underlying data generating processes. Thus, it is very important to understand how an emerging data is generated and identify appropriate methods for the data.
- *Missing trips.* Missing trips is a common problem for both datasets, though for different reasons. Because emerging datasets are not generated to capture users' complete activity and travel patterns as household travel surveys are designed to do, it is unavoidable that certain trips will be missed. However, different types of trips are likely to be missed in different datasets. It is thus important to investigate the kinds of trips that are missed and develop ways to compensate them.
- *Activity durations and trip times.* Because of the missing trips, the estimated activity durations and trip times will not be accurate. And depending on the kinds of the trips missed, activity durations and trip times of certain trips can be over- or under-estimated. This says again that prior to the calculation of activity durations and trip times, there is a need to compensate those missing trips first.
- *Mode detection.* Due to the sparsity associated with the mobile phone data, mode detection will pose significant challenges if not impossible.
- *Presence of zero cells.* If an emerging dataset is used, a substantial portion of zones in a region will have zero cells, or cells with no trips originating from or destining to. This problem will be mitigated when a longer series of data is used, but will not disappear. Methods should be developed to address the zero cell issue.
- *Upscaling to the regional total.* Because of the missing trips, the estimated OD trips from the emerging datasets must be upscaled such that the total equals to the regional total estimated from MPOs. The existing literature has a variety of methods for the upscaling and the corresponding upscaling factors are very different. It is not hard to imagine that these varying upscaling factors lead to very different resulting OD matrices. Certainly much more research is needed in this area.

2.0 Introduction

2.1 Disclaimer

The views expressed in this document do not represent the opinions of FHWA and do not constitute an endorsement, recommendation or specification by FHWA. The document is based solely on the research conducted by the University of Washington.

2.2 Acknowledgments

The FHWA would like to acknowledge the assistance of two metropolitan planning organizations (MPOs) who generously agreed to share their models and provide some of their time for this study: the Greater Buffalo Niagara Regional Transportation Council and the Puget Sound Regional Council.

2.3 Emerging Data Sources for OD Estimation

Origin-Destination (OD) estimation for a metropolitan region is a critical component in the regional transportation planning process as its results (OD tables) not only paint a realistic picture of how the region's population travels across time and space but also forms the foundation on which a wide range of decisions are made to improve the efficiency, safety and reliability of the transportation system. OD estimation results can also be used to model the unintended consequences (e.g., pollution) caused by interventions into a large-scale, modern transportation system.

OD estimation requires time- and location-stamped data from which individual movement patterns can be inferred. Aside from traditional data sources such as household travel surveys and license plate surveys, an increasing list of datasets emerge due to the prevalence of devices and services that can now provide information on when and where a user is.

In this report, emerging datasets are broadly defined as those of passive solicitation, meaning that such data is not generated for transportation applications (or more specifically in the context of this project, for OD estimation), but is a by-product of other non-transportation-related purposes (Chen et al. 2016). The following provides some illustrations of such emerging datasets:

1. Anonymously generated mobile phone data: the primary purpose is for billing and operation by mobile phone companies.
2. Global positioning system (GPS) data: the primary purpose is to provide navigation or other types of location-based services to users.
3. Social media data: the primary purpose is to provide users a platform for their social interactions.

Table 1. Types of data and data providers in the United States.

Types of data	Data providers	Source of data product	Product examples
Mobile phone data	AirSage (airsage.com)	Mobile phone data	OD trip matrices for specified geographic areas OD commuting (home–work) trip matrices for specified geographic areas Route-based OD matrices for specified geographic areas
Mobile phone data	StreetLight (streetlightdata.com)	Location-based service data, mobile devices' pings	Providing travel data through web browser Origin/Destination Matrices, Select Link Analyses, Average Travel Times and Travel Time Distribution, Internal/External Studies, and Commercial and Personal Travel Vehicle Comparisons. *Metrics can be customized to specific time of day, day of week, time of year
GPS data	TomTom (tomtom.com)	TomTom navigation device GPS data, GSM data, traditional road agency data	in-vehicle navigation devices in-dashboard navigation and vehicle control services navigation software for mobile devices.
GPS data	INRIX (inrix.com)	Mobile devices	Trips reports: aggregated lists of anonymized trips with route details Trip matrices: OD matrices Raw trips records and matrices Traffic data: bottleneck ranking, congestion area
GPS data	Google (google.com)	Crowd-sourced GPS data collected from smartphones	GPS navigation, Traffic alerts Transit directions, etc. Provide: travel time, travel distance, place of interest
GPS data	TrafficCast (trafficcast.com)	GPS tracking data, public road sensors, incident/accident reports, construction & work zones, weather and major community events	DYNAFLOW (Dynamic traffic flow): provide real-time and predictive traffic information, supporting route planning, personal and enterprise travel times, and other transportation management services

This report focuses on the first two of the above-listed datasets: anonymously generated mobile phone data and GPS data. A summary of data sources, providers, and currently available product examples can be found in Table 1.

The rest of this chapter provides an overview on how mobile phone and GPS data are generated and have been used for transportation planning purposes, followed by an analysis framework that

guides the investigation of the two datasets (Chapters 4 and 5). Chapter 3 provides the state of the art review of the existing studies using mobile phone and GPS data, in particular, activity or trip end identification, mode detection and OD analysis. A discussion of the issues identified in the existing studies is also presented in Chapter 3. Chapters 4 and 5 provide the detailed analysis results for mobile phone and GPS data, organized by zeroth, first, and second properties, corresponding to characteristics associated with the data, a single activity location or trip end and trips (two activity locations or trip ends). Chapter 6 presents a discussion on the implications of the results from Chapters 4 and 5 on OD estimation and broader transportation planning analysis. Methodologies used in deriving trip ends and trips are provided in the appendix.

2.4 Mobile Phone Data

Earlier applications of mobile phone data for transportation research can be traced back more than two decades when cellular tower data were used for travel time estimation in (Yim & Cayford, 2001). The first significant study that used a large amount of mobile phone data for analyzing mobility patterns was conducted about 10 years ago (González et al. 2008). The most important result of that study is on the regularity of human mobility patterns, observing the short travel distances and high levels of frequency to visit a limited number of activity locations for the vast majority of the people, and thus leading to the conclusion that human travel patterns exhibit a high level of regularity. Since then there has been an explosion of such studies. A search in the literature using keyword combinations of mobile phone data, mobility, and travel behavior resulted in more than 1,000 articles published in journals across different disciplines. These articles cover a wide range of topics including: estimating mobility patterns (Csáji et al. 2013; González et al. 2008; Song et al. 2010; Chen et al. 2014), inferring OD metrics (Calabrese et al. 2011; Iqbal et al. 2014), finding anchor locations (Dong et al. 2015), inferring activity types and patterns (Jiang et al. 2017; Widhalm et al. 2015a) and travel modes (Qu et al. 2015; Wang et al. 2010).

Cellular tower networks generate data from phone calls, messaging and other activities requiring network connectivity, and this data includes the locations of mobile devices through time. When a mobile device connects to cell towers, the distance between the mobile phone and the interacted cell towers is estimated with several signaling-related records, including received strength of signals and angles from the towers, and signaling duration between the towers and the mobile device. In the past, the locations of mobile devices were estimated with a limited amount of information such as the location of cell towers and their service areas (Skyhook Wireless, 2008; Zandbergen, 2009). Accuracy of the location estimates has been improved now with additional information such as geometric techniques (triangulation algorithms) and statistical methods (Gezici, 2008). With the geometric techniques, the position of a mobile phone is estimated as the intersection of position lines obtained from the cell towers' locations and the received strength of signals or signaling durations. Statistical techniques utilize the signaling information between mobile devices and cell towers, with their parameters to estimate position of mobile devices.

The accuracy of location positioning is heavily dependent upon the density of cell towers regardless of positioning algorithms. It is generally reported that the error range of cellular positioning is 50 to several hundred meters in urban areas, and several hundred meters to several kilometers in rural areas (Lin & Juang, 2005; Mohr, Edwards, & McCarthy, 2008; Weiss, 2003; Zandbergen, 2009). Additionally, the quality of devices, radio frequency, device user, and cellular

carrier coverage are also found to be important factors influencing the accuracy of cellular positioning (Hard et al. 2016).

Mobile phone data can usually be categorized into two types, based on the types of records included to generate mobile phone dataset. One type is call details records data, and another is called "sightings data." The former only records phone calls with locations of the towers that channel the calls, and the latter is generated each time a mobile phone is positioned regardless of types of signals (Calabrese et al. 2013; Chen et al. 2016). These two types of datasets have different characteristics, and the detailed discussion about their advantages and disadvantages in the mobility analysis and transportation applications can be found in Chen et al. (2016, p288). This report focuses on the ***second type of mobile phone data: sightings data***. Unless otherwise specified, we use "***mobile phone data***" throughout the report to refer to sightings data. Such data has now generated many papers in the field (Calabrese et al. 2011; Calabrese et al. 2013; Jiang et al. 2013), and is already being used by several planning agencies and departments of transportation in the US. A number of commercial vendors are now widely advertising products (e.g., OD matrices) derived from such data (Liu, Danczyk, Brewer, & Starr, 2008). Thus, having a clear understanding of such data and the techniques that can be applied to process such data is not only important for researchers but also for practitioners who may be interested in using such data or their derived products.

2.5 GPS Data

GPS was originally designed for military and intelligence applications in the 1960s and was not opened to the public until 1980s. GPS satellites broadcast radio signals providing their locations, status, and time from onboard atomic clocks. The signals travel through space and are received by GPS receivers with their exact arrival times. Once a GPS receiver can view at least four satellites (Zandbergen, 2009), based on the time difference, geometric techniques can be utilized to locate the receiver's position on Earth in three dimensions (GPS.gov, 2013). Even with some errors due to inaccurate time-keeping by the receiver's clock, GPS data usually has the highest accuracy and precision levels compared to other types of signals such as tower triangulation. GPS precision is often reported to be in the range of 1 to 10 meters (van Diggelen and Enge, 2015). This level of precision creates superior advantages over other location methods in locating vehicle origin and destination in zones, as well as the traveling routes.

However, frequent GPS signal processing consumes a large amount of power. In addition, the accuracy of GPS data highly depends on signal transmission and geometric techniques. The strength of GPS signal is affected by weather conditions, and lack of direct vision to the satellites prevents receivers from getting accurate signals. For example, the urban canyon issue of GPS data has been well recognized and studied in the past (Walsh et al. 1997; Phatak et al. 1999; Jang et al. 2000; Ray et al. 2001; Cui & Ge, 2003; Stopher & Greaves, 2007; Stopher et al. 2008a; Chen et al. 2010; Gong et al. 2012).

One of the early uses of GPS data in transportation was probe vehicles (e.g., floating cars), to collect traffic/transportation data for various purposes (Sanwal and Walrand, 1995; Rakha and Van Aerde, 1995; Barth et al. 1996), e.g., evaluation of traffic signal performance or corridor travel times (Ban et al. 2014). A prominent example of this was the HICOMP program of Caltrans (Schwarzenegger et al. 2008). GPS data collected via dedicated probe vehicles are typically

limited in terms of spatial and temporal coverage. On the other hand, the data collection design was often carefully conducted to make sure the data could be used effectively.

Currently, GPS data primarily come from four sources: 1) targeted survey studies that use GPS devices or leverage mobile phones for GPS signals (e.g., household travel surveys with GPS components), 2) naturalistic driving studies (Dingus et al. 2006), 3) data from navigation devices in fleets (such as trucks, taxis, and on-demand service providers) and cars, and 4) mobile apps. The first two types of data are based on active solicitation (Chen et al. 2016), meaning that the underlying data generation process for such data is carefully designed with a defined purpose for transportation research and applications. The latter two types of data are of passive solicitation (Chen et al. 2016), meaning that data is not generated from a well-defined process. In the future, when connected and autonomous vehicles are widely deployed, they will help generate a large amount of GPS data. This future scenario is not further discussed in this report.

For targeted travel survey studies that used GPS devices, the earliest study is the 1996 Lexington study funded by the US Department of Transportation (Murakami & Wagner 1999). Since then, there have been quite a number of regional GPS-based travel surveys implemented (Wolf, 2006; Kochan et al. 2006; Murakami et al. 2004; Kunzmann et al. 2013). Except for the GPS component providing detailed and accurate trajectory information, such studies are in principle no different from traditional household travel surveys without the GPS component: they all contain rich behavioral information but are limited in sample size. The second type—naturalistic driving studies—with its primary purpose of enhancing safety, is designed to provide rich information on driving behavior, vehicle behavior and interaction with the environment in which the vehicle is traveling. Travel-related information (such as OD information) might be derived from such data, although such studies are currently sparse in the literature.

Today, an increasing number of vehicles (trucks and cars) are equipped with GPS through their navigation devices. Thus, a large portion of GPS data comes from navigation devices (the third type noted above). Additionally, there is also emerging GPS data from mobile phone apps. These two types constitute the majority of the GPS Big Data concerned in this report: passively generated and often big in size. In particular, the GPS data studied in this report were from in-vehicle navigation devices or mobile apps, thus mainly representing vehicular trips. Such data is the basis of a larger number of studies for traffic state monitoring and estimation, including traffic speed (Tao et al. 2012; Work et al. 2008), traffic density (Herrera & Bayen, 2010), travel time (Ban et al. 2009; Hunter et al. 2009; Zhan et al. 2013), queue lengths (Ban et al. 2011; Hao et al. 2014, 2015), urban freight delivery performances (Yang et al., 2014), and urban traffic signal performance (Hao et al. 2012; Sun et al. 2015). Additionally, when combined with other wireless communication technologies, such as cellular network, microwave communication, and Wi-Fi, GPS enabled devices show their great potential to monitor urban space patterns (Liu et al. 2012; Shoval, 2008) and mobility patterns (Liu at al., 2009; Su et al. 2000; Peng et al. 2012).

Like mobile phone data vendors, GPS data vendors collect, purchase, and integrate GPS data from various sources (i.e., the first-hand GPS data), remove sensitive information, possibly integrate with data from other sources, and then sell the data to government agencies or to businesses. The GPS data investigated in this report are those processed data from the vendors, with the aim of providing as much information as possible on the nature of such data.

2.6 Analysis Framework

The purpose of this study is to understand the nature of the two types of datasets—mobile phone and GPS data—their various properties, and how they compare against existing traditional data sources (e.g., MPO model results) for the same geographical coverage.

The analysis framework for this study (Table 2) is divided into three categories of data properties: zeroth order, which describes various aspects of the dataset; first order, which describes properties related to single activity locations, or trip ends; and second order, which illustrates properties related to trips (two linked activity locations or trip ends). Note that not all properties are equally applicable to both datasets, as the two differ from each other in several aspects. When a property is applicable to one dataset only, it is specifically noted in the "definition" column in Table 2. The nature of the two datasets (mobile phone and GPS data) is described in Section 4.0 and Section 5.0 respectively and the comparisons of their properties is discussed in Section 6.0.

Table 2. Analysis framework: zeroth-, first-, and second-order properties.

Order	Properties	Definition	Variations
Zeroth Order	Study period	Time period observations are available for study (90 days for GPS data; 30 days for mobile phone data)	N/A
Zeroth Order	Number of unique IDs in study period	Counts of the unique IDs	Daily; weekly; monthly
Zeroth Order	Sampling interval	GPS: distribution of the observation intervals (between two consecutive observations) Mobile phone: 1) distribution of intervals (between two consecutive observations); 2) daily temporal sparsity (by dividing a day into 24 hourly intervals, the number of intervals with at least one observation)	Daily, Weekly and monthly; Weekdays vs weekends
Zeroth Order	Pattern of observations	The number of observations over different times of a day	Daily; weekly; Weekdays vs weekends
Zeroth Order	Number of day observed	Number of days a unique ID was observed during the study period	N/A
Zeroth Order	Lifespan of unique IDs	Time difference between the first time/day and the last time/day that a unique ID was observed	N/A
Zeroth Order	Spatial distribution of the observations	Total number of observations in different zones	Weekdays vs weekends
First Order	Location accuracy	Mobile phone data: accuracy of location estimates provided by the vendor and this investigation (Not applicable for GPS)	N/A
First Order	Location uncertainty	Mobile phone data: extent of the location estimates representing the actual location	N/A

Order	Properties	Definition	Variations
		(*Not applicable for GPS*)	
First Order	Oscillation ratio[1]	Mobile phone data: The proportion of number of observations generated due to oscillation (*Not applicable for GPS*)	N/A
First Order	Activity Duration	GPS data: 'Stop time' between two consecutive trips	Daily, weekly, monthly; Weekdays vs weekends
First Order	Activity Duration	Mobile phone data: difference in observed times at an activity location	Daily, weekly, monthly; Weekdays vs weekends
First Order	Spatial distribution of zone-level trips origins from a zone	Spatial distribution of zone-level trips origins from a zone	Weekdays vs weekends; daily variations
First Order	Correlation of zone-level trip origins with the population/MPO results	Correlation of zone-level trip origins with the population/MPO results	Weekdays vs weekends; daily variations
First Order	Spatial distribution of zone-level trips destinations from a zone	Spatial distribution of zone-level trips destinations from a zone	Weekdays vs weekends; daily variations
First Order	Correlation of zone-level trip destinations with the population/MPO results	Correlation of zone-level trip destinations with the population/MPO results	Weekdays vs weekends; daily variations
Second Order	Distribution of trip rates	Distribution of trip rates per day per user	Weekdays vs weekends; Daily variations
Second Order	Distribution of departure/arrival times	Distribution of departure/arrival times of trips	Time of the day; weekdays vs weekends; daily and weekly
Second Order	Distribution of trip times	Distribution of trip times	Weekdays vs weekends; daily variation
Second Order	Percentage of zero cells (trip origins, destinations, and origins and destinations)	Percentage of cells that have zero trips originating from (trip origins), destining to (destinations) a particular cell (TAZ); percentage of cells that have zero trips originating from and destining to a particular cell (trip origins and destinations)	Daily variations; trends with different accumulation of data; Weekdays vs weekends
Second Order	Spatial distribution of cells with no trips 1) generated/ 2) attracted	Spatial distribution of cells which no trips originated from or destined to	N/A

[1] Oscillation is a data issue that records are generated due to pure signaling activities of devices and do not reflect movements of devise holders. More discussion can be found in Section 4.2.3.

Order	Properties	Definition	Variations
Second Order	OD collinearity	Correlation between observed number of trips and estimated OD (after upscaling)	N/A
Second Order	OD compared with MPO model results	Comparison between estimated OD with PSRC/Buffalo MPO model results	N/A
Second Order	OD Sensitivity Analysis	Sensitivity analysis on how different parameters used in OD estimation affect the OD calculation	N/A

3.0 Literature Review

3.1 OD Methods

The origin-destination (OD) estimation methods from mobile phone data and GPS data usually consist of four major steps:

1. Activity location (trip end) identification—activity location estimation is required to differentiate origins and destinations of trips from locations identified in both passively generated mobile phone data and GPS data.
2. Mode detection—the second step is to infer travel mode from the data.
3. Zone-level trip aggregation—The third step is to aggregate individual trip information into zone-level (such as traffic analysis zone (TAZ) or census tract).
4. Upscaling to OD demand matrices—to upscale the observed zone-level trip information to an estimated OD trip table so that the total number of trips in a region is the same between the two tables (e.g., from mobile phone data vs from MPO model results).

Travel routes may also be detected from mobile phone data or GPS data, which is not strictly required for OD estimation purposes. For the completeness of this report, a review of route detection using the two types of data is also provided in A.3 Literature Review of Route Detection.

While these four steps represent a general OD estimation procedure using mobile phone or GPS data, large variations exist for methods that are based on different data sources. In particular, the third step on aggregating individual trips to zone-level trip information may be challenging if mobile phone data are used due to its low location accuracy. This, however, may be more easily done based on map-matching techniques if GPS data are used.

In addition, the methods to identify trip ends using the two data sources can be different due to their distinct temporal and spatial resolutions. Usually, GPS data have higher spatial positioning accuracy and temporal resolutions than mobile phone data. Trip end identification or mode detection may be done for a single vehicle trajectory if GPS data are used. However, if mobile phone data are used, because the locational information is provided as cell tower IDs or the triangulated locations, trip end identification often needs to be done by accumulating observations from multiple observation periods for the same person (or device) (Bonnel et al, 2015; Iqbal et al. 2014a; Larijani et al. 2015). On the other hand, important trip generation information such as home/work locations may be more easily derived from using mobile phone data than using GPS data.

The number of trips extracted from the first three steps is the *observed* trips from the data, which is often (much) smaller than the *actual* number of OD trips. Therefore, the fourth, upscaling step is to expand the observed trips to the actual OD trips. Usually certain upscaling factor (also called the expansion factor) is applied. The factor is mostly calculated based on the comparison between the number of users/vehicles in the mobile phone or GPS dataset and some ground truth data such as census population, traffic counts, or OD tables from MPO models (which are calibrated using both survey data and traffic counts). In the rest of this chapter, a review of each step is presented. Each review is organized by the methods based on mobile phone data and GPS data,

with a brief discussion of the similarities and differences if necessary. The summaries of different applications and OD methods are described in Table 4 and Table 5.

3.1.1 Activity Location (Trip End) Identification

Mobile Phone Data

In mobile phone data, multiple unique locations may be recorded even when a user stays at a single place, especially for sightings data (which is the subject of this study). As the location estimation is the result of triangulation among multiple towers (Chen et al. 2016; Widhalm et al. 2015), each of them is unique. In other words, even when a user stays at the same location, locations recorded in the data differ from time to time (though they may scatter in close proximity). The common way to reveal actual locations where users have stayed is to aggregate these location records by applying a clustering algorithm (Calabrese et al. 2011; Chen et al. 2016; Jiang et al. 2013). Typically, the centroid of the outputting cluster is used to represent the actual location.

Conventional clustering methods are usually not applicable mainly because parameters in these methods need to be predetermined, which is infeasible in mobile phone data. For example, popular methods such as k-means and Density-based Spatial Clustering (DBSCAN) (Ester et al. 1996; Kanungo et al. 2002) require the number of clusters or density-related parameters be known, which cannot be identified because of the variety of mobile phone usage and travel behaviors that exist across individuals (Chen et al. 2014).

Chen et al. (2014) suggested a model-based clustering method. Rather than relying on a predetermined threshold value for all devices, this method determines the optimal number of clusters for each device through a statistical model by searching through a finite set of possible numbers of clusters and finding the one that generates the highest Bayesian Information Criterion (Fraley and Raftery, 2002). The centroid of an identified activity cluster (defined as a sequence of traces that are considered to be activity places) is used as activity locations for the subsequent modeling processes. The advantage of applying the model-based clustering is obvious: there is no requirement for predetermined clustering parameters.

More recently, a more advanced clustering method to aggregate traces is the incremental clustering method (Alexander et al. 2015; Hariharan and Toyama, 2004; Wang et al. 2015; Widhalm et al. 2015; Wang and Chen, 2017). Given traces $\{d_0, d_1, d_2, \ldots, d_k\}$ in one trajectory, the clustering goes in four steps: 1) starting from trace d_0, one new cluster C_0 is created and d_0 is the center; 2) each trace that is not clustered will be checked and the trace within a distance R_c to the center of C_0 is aggregated to the cluster; 3) every time the cluster grows, its center is updated; 4) if no trace could be aggregated in the current cluster, one new cluster will be created containing a nonclustered trace. This procedure repeats itself until all traces are clustered. After the clustering process, clusters with a time duration no less than T_C will be considered as activity locations. This clustering method is shown robust to outliers present among traces, as the mechanism of updating centers (step 3) improves its tolerance to outliers (Widhalm et al. 2015).

GPS Data

GPS data contain a continuous stream of trajectory information. Finding the trip ends means splitting each continuous trajectory into smaller segments whose ends are where activities are conducted. A **stop** (or **stay**) is identified when a person or vehicle stays within a certain area for

some time. Stops usually can be categorized into two kinds: **activity stops** and **nonactivity stops**. Activity stops refer to those stops intended for activities (e.g., working at work places, shopping at malls or grocery stores, and studying at school), while nonactivity stops are often not affiliated with an activity purpose (e.g., waiting for green light at intersections, being stuck in congested traffic). Since a trip is always affiliated with some purposes and is usually followed by activities, this report treats the locations of activity stops as trip ends. Identifying stops and distinguishing these two kinds of stops in a trajectory are the essential tasks in the overall analysis of trip end identification using GPS data.

Researchers have developed different approaches to identify trip ends from GPS data. The methods can be grouped into two broad categories: threshold-based methods and machine learning based methods. Threshold-based methods are usually applied in density-based spatial clustering (Palma et al. 2008; Zimmermann et al. 2009; Kami et al. 2010; Tran et al. 2011; Gong et al. 2015), which are probably the most widely used methods for identifying trip ends from GPS traces. Duration (dwell time) is one of the prime parameters used in such methods. Different values have been considered as the threshold dwell time, ranging from 45 to 900 seconds, such as 45, 120, 180, 200, or even 900 seconds (Stopher et al. 2008a; Schuessler and Axhausen 2009; Gong et al. 2012 etc.). These values mainly depend on the local transportation situation or characteristics of the GPS data. Some researchers (Wolf et al. 2000; Draijer et al. 2000; Wagner, 1997; Wagner et al. 1998; Mizuno et al. 2013) also considered the instantaneous speeds in the methods. They regarded zero speeds as the necessary condition for a trip end. In addition, the change of heading and the density of the GPS points were also considered in some studies (de Jong and Mensonides, 2003; Du and Aultman-Hall, 2007). The issue with the density-based methods is that these methods require data to be collected at more frequent intervals (Gong et al. 2015).

Other researchers applied machine learning based approaches, such as the random forest models (Zhou et al. 2016) and support vector machines (SVM) (Mizuno et al. 2013; Yang et al. 2014; Gong et al. 2015). Yang et al. (2014) used an SVM-based classifier for urban delivery stop identification of freight delivery trucks using GPS data, with high identification accuracy (94–99%). They used three features to distinguish delivery stops and nondelivery stops: stop durations, distance between stop location to the center of the city, a stop's binary distance to the closest major bottleneck (such as bridges and tunnels). Gong et al. (2015) developed an SVM-based model to detect trip ends from GPS data and achieved an accuracy of 94%. The features they used in the model to distinguish between activity stops and nonactivity stops include stop durations, distance between a stop location and the home or workplace, and the mean distance of GPS points belonging to the cluster centroid. Zhou et al. (2016) used a Random Forest Method for trip ends identification. They considered 24 different features from GPS tracking points, which includes: (1) local attributes, like time difference between two consecutive records, instantaneous speeds, etc. (2) global extreme value attributes, such as total distance, and travel times, etc., (3) speed-related attributes, like average speeds and standard deviation instantaneous speeds, (4) acceleration-related attributes, such as average absolute acceleration, largest absolute acceleration, (5) tracking points clustering attributes, like the largest distance between any two points within one neighborhood point set, density of the points in each neighborhood point set, and (6) heading change attributes like average heading change in the neighborhood point set.

Comparison between mobile phone and GPS data. The trip end identification methods for the two data sources have some similarities (e.g., many of them applied machine learning methods to best utilize the available data). While some used predefined thresholds in the methods, others relied more on the learning methods to derive these parameters automatically (e.g., from the training step). The major differences are the actual parameters and features that were applied in the methods, largely due to the different temporal and spatial properties of the two data sources.

3.1.2 Mode Detection

Mobile Phone Data

Few studies have attempted to infer mode choices from the passively-collected mobile phone data. Prior to inferring mode choices, points obtained from the mobile phone data are first overlaid with the network data (e.g., roadway or transit networks) as is in the case for route choices. Afterwards, one can identify, for example, air travel, by geo-referencing trip ends around airport locations within a certain range (e.g., 3 miles) and with certain trip durations (Ma et al, 2012). An alternative, more direct method is to apply the unsupervised *k*-means clustering algorithm to partition the records (in the mobile phone data) into car or transit travel by their travel speeds, if these points fall within the same origin and destination areas (Wang et al. 2010).

Some studies have actively collected cell tower signals through mobile phones, in which case the estimated speed of a mobile phone, the fluctuations of signal strength, and the change rates of connected cell towers are key information to detect travel modes (Anderson & Muller, 2006; Reddy et al. 2010; Welbourne, Lester, LaMarca, & Borriello, 2005). Passing-by locations are also important to detect the travel mode. For example, if passing-by locations of a trip are continuously positioned on or close to roads identified on a road network with a speed higher than 20 mph, this trip may be classified as a motorized trip. These methods are similar to those used for mode detection from GPS data, as the traces recorded are continuous and dense and the difference between the two is that cell tower signals are likely much less accurate than GPS signals. In general, the relatively low accuracy and thus only limited mode types that can be identified are recognized issues among studies using cell tower signals for mode detection (Anderson and Muller, 2006; Reddy et al. 2010).

GPS Data

In addition to inferring trip-end information based on GPS traces, GPS data can also be used to detect transportation modes. GPS-based mode detection methods can be grouped into three categories: machine learning methods, probability methods, and criteria-based methods (Gong et al. 2014). Machine learning methods are attracting more attention in the field of travel mode detection and include many specific techniques such as: conditional random field (Zheng et al. 2008; Zhang et al. 2011), SVM (Zheng et al. 2008), Bayesian network (Zheng et al. 2008; Feng et al. 2013; Xiao et al. 2015), decision tree (Zheng et al. 2008), and multilayer perception (González et al. 2008). Probability methods usually contain two specific techniques: fuzzy logic rules (Schussler and Axhausen 2009; Xu et al. 2010) and probability matrix (Stopher et al. 2008a). Methods in the third category set several criteria for a variety of attributes of movement including speeds, duration/time, distance, acceleration, and heading to detect modes. Such methods have been utilized by many researchers for a long time and often involve other types of data such as GIS information (Bohte et al. 2009; Chen et al. 2010; Stenneth et al. 2011; Gong et al. 2012). By

combining methods from the above three categories, more accurate mode detection results may be achieved. For example, by combining the rule- (or criteria-) based classifiers with Hidden Markov models (Widhalm et al. 2012), eight modes (bus, car, bike, tram, train, subway, walk, and motorcycle) can be distinguished and identified from GPS data. Nitsche et al. (2014) combined the probabilistic classifiers with a Discrete Hidden Markov Model that allowed them to detect nine different modes including walk, bicycle, motorcycle, car, bus, electric tramway, metro, train, and waiting.

As a summary, while data analytics methods and tools have been generally applied to both mobile phone data and GPS data for mode detection, more accurate mode information can be detected using GPS data based on a variety of methods. In this project, the mode detection step is not implemented for either of the two data sources. In other words, the focus is on the total travel demands; and for the GPS data, it is mainly for the total vehicle-based trips.

3.1.3 Zone-level Trip Information Aggregation

After identifying the trip ends and mode information for individual trips, the next step is to aggregate such individual trip information into zones to obtain the (observed) OD trip information. Depending on the specific applications, zones can be TAZs or census tracts. If mobile phone data are used, different approaches can be applied for zone-level trip aggregation, depending on the types of mobile phone data sources. If locational information is provided in the form of cell towers, Voronoi polygons can be calculated and are then connected to the nearest road network nodes or metro stations, which can be used to map individual trips expressed in a sequence of specific zones (Iqbal et al. 2014; Larijani et al. 2015; Bonnel et al. 2015). However, if the locational information is provided as triangulated and processed latitude and longitude, trip information must first be aggregated to a predetermined zone structure, such as grids, census tracts, or TAZs (Calabrese et al. 2011; Alexander et al. 2015). It was found that using finer geographical areas for aggregation could cause noisy and unbalanced OD representation (Zandbergen, 2009), and a recent study found that it can be alleviated when the zone size becomes larger (Alexander et al. 2015).

If GPS data are used, since the data contain accurate location information, individual trip information such as trip ends can be readily aggregated into zones to obtain zone-level trip information. In practice, it is possible that the trip ends happen to be on the boundary of two zones. In this case, certain rules/procedures are needed to break the tie.

3.1.4 OD Estimation with Upscaling

Zone-level trip aggregation produces only the observed OD trips from the data. To obtain the estimated total trips between an OD pair (i.e., the trip table), the observed trips need to be upscaled so that the total number of estimated trips from an origin to a destination also accounts for unobserved trips. Several existing upscaling methods were applied to mobile phone data.

One way to compute the upscaling factor is to compare the observed OD table from mobile phone data with traffic count data. With this method, the upscaling factor calculation involves the traffic assignment step, and can be computed either iteratively (Ma et al. 2013) or formulated as an optimization problem (Iqbal et al. 2014; Toole et al. 2015). Ma et al. (2013) also applied the fuzzy logic based method to impute missing trips between each individual's activity locations, and

upscaled the derived trips using the CTPP (Census Transportation Planning Package) data (Ma et al. 2013).

Another way to calculate the scaling factor is using the ratio of census population counts and the number of observed mobile phone users (Alexander et al. 2015; Calabrese et al. 2011; Wang et al. 2012). In this case, the criteria and the procedure used to infer home locations from the mobile phone data becomes critically important because the scaling factor is the ratio between the population counts and the number of inferred home locations from the mobile phone data (Alexander et al. 2015; Çolak et al. 2015).

Although both Wang et al. (2012) and Alexander et al. (2015) used call-detail-record (CDR) data[2] as the input, there are several significant differences in their approaches. While Wang et al. (2012) emphasized the vehicle trips and different groups of mobile users based on their mobile phone usage, Alexander et al. (2015) focused on stay extraction and inferring users' home locations. The ratio of the census population and the number of residents identified in the CDR data was used as the scaling factor in the method proposed by Alexander et al. (2015), where Wang et al. (2012) applied two additional parameters: the hourly trip production for the entire population and the vehicle usage ratio (VUR) to upscale the observed vehicle trips. Unlike Alexander et al. (2015), Wang et al. (2012) included the VUR to consider the effects of different transportation modes in their study. On the other hand, Alexander et al. (2015) focused on activity types (e.g. home, work, or other) that were not considered in Wang et al. (2012). Differences also exist in the validation processes of these two studies. While Wang et al. (2012) compared the predicted travel time for each road segment with the average travel time calculated from the probe vehicle GPS data, Alexander et al. (2015) compared their results with the 1991 Boston Household Travel Survey and the 2010/2011 Massachusetts Travel Survey. A summary of the comparison of the two upscaling methods is provided in Table 3.

Recently, the preprocessed data from commercial data providers, such as AirSage, StreetLight, and INRIX, are now starting to be used by many state departments of transportation in the US and Metropolitan Planning Organizations (MPOs) to validate and calibrate OD matrices from traditional models, including North Carolina Department of Transportation, Syracuse, Idaho, etc. (Bernardin et al. 2017; Bindra, 2016; Cambridge Systematics, 2011; Fussell, Gresham, & Smith, 2013; Huntsinger & Donnelly, 2014; Huntsinger & Ward, 2015; Milone, 2015; RSG 2015). Although detailed methods are not disclosed, the providers claimed that several factors (e.g. data penetration rate and population census data) are used to upscale the mobile phone based OD matrix (Hard et al. 2016).

[2] CDR data are based on phone calls only where each record contains information associated with a phone call.

Table 3. Comparison of methods proposed by Wang et al. (2012) and Alexander et al. (2015).

Attributes	Method proposed by Wang et al. (2012)	Method proposed by Alexander et al. (2015)
Study Area	San Francisco Bay Area and Boston	Boston metropolitan area
Number of Users	360,000 (Bay Area) 680,000 (Boston)	2,000,000
Coordinates of the Records	Cell tower ID (Bay Area) Triangulation (Boston)	Standard triangulation algorithm
Study Period	Three weeks	Two months
Stay Extraction	No	Yes
Consider User's Home Location	No	Yes
Consider Different Groups of Mobile Users	Yes	No
Considering Activity type	No	Yes
Considering Vehicle Trips	Yes	No
Upscaling Factor	Hourly trip production for the entire population	The ratio of the 2010 Census population and the number of residents identified in the CDR data
Validation	Using travel times from probe vehicle GPS data	Using traffic flows from local survey data

While sparse published results were found for the upscaling step using GPS data, it is assumed in this project that similar methods, as presented above in this subsection, may also be applicable to GPS data. Certain challenges may exist when using GPS data for upscaling, such as identifying the home locations of travelers/vehicles. This is especially true if the vehicle/device ID changes frequently (e.g., every few hours) or home-related trace information is not available (or filtered purposely for privacy considerations) in the GPS data. Certain assumptions have to be made when the upscaling method is applied to GPS data, as discussed in more detail in Section 5.0.

3.2 Issues Identified

3.2.1 Mobile Phone Data

Low Accuracy of Cellular Positioning

As discussed, several factors affect the accuracy of cellular positioning, and this indicates the accuracy levels of the estimated activity locations and extracted trips are different between users and across different areas (Hard et al. 2016). Therefore, this issue is related to all types of applications using mobile phone data. In the traffic state monitoring applications, it was generally reported that the low level of location accuracy produces more errors at the AM and PM peak hours or in congestion (Liu et al 2008; Maerivoet and Logghe 2007; Thiessenhusen et al. 2003, 2006). This presumably happens when there is a high density of mobile devices located in close proximity, like an arena. In addition, this issue is related to the extraction of short-distance trips because it is difficult to distinguish the short-distance trips due to location positioning errors. In OD estimation, it is often suggested to use large zone systems with spatial aggregation rather than small zone systems because larger zones reduce the low accuracy issue related to cellular positioning (Alexander et al. 2015; Çolak et al 2015; Calabrese et al. 2013; Milone, 2015; Chattanooga 2011; Fussell et al. 2013; Huntsinger and Donnelly 2014; Huntsinger and Ward

2015; Bindra 2016; RSG 2015; Bernardin et al. 2017). Moreover, the accuracy of route and mode estimations can be low in areas with a dense road network because there exist many alternative route possibilities (Tettamanti et al. 2012).

Self-selection

Self-selection results from an individual user's self-selecting into using certain services (e.g., subscribing to a specific phone company) or belonging to certain groups (e.g., particular social media groups), often with their devices. Consequently, the dataset that results as a side product of some primary operational purposes comprise users using one or more specific services or belonging to certain groups. Unless everyone in a region has an equal probability of using one service or belonging to a particular group, such datasets cannot be representative of the people living in that region. While it is possible that large sample size may be associated with a higher level of representativeness, there is currently little to no understanding about how increasing sample sizes may address representativeness related issues. On the top of the user selection issue (whether a particular user is part of the dataset or not), mobile phone data is typically characterized by its temporal sparsity,[3] which means that they often lack enough information for a full-spectrum mobility pattern extraction. Existing studies have applied different user selection criteria to identify a subset of users from a mobile phone dataset (Lu et al. 2017; Widhalm et al. 2015; Zhao et al. 2016) or developed scaling factors to account for those users or trips that are not observed directly in the dataset (Alexander et al. 2015; Çolak et al. 2015; Iqbal et al. 2014). Little knowledge is available on the effectiveness of these methods and addressing self-selection and the associated representativeness issue is a critical area of research in the near future.

Lack of Sufficient Details

The mobile phone data often comes only with spatio-temporal information from which a device's movement pattern can be inferred. Other trip related information essential for OD estimation such as trip purposes, types of trips (Calabrese et al. 2013; Yin et al. 2017), mode and route choices must be inferred. Though various approaches have been developed as described in this chapter, little knowledge has accumulated on the validity of the approaches and their effectiveness in representing the underlying population (Chen et al. 2016). In fact, there is much unknown about the underlying population.

3.2.2 GPS Data

Signal Blockage When Using GPS Data

One concerning problem that often arises when using GPS devices is signal loss or serious degradation of the signal in certain circumstances due to the blockage of GPS signals (Stopher & Greaves, 2007; Stopher et al. 2008b; Gong et al. 2012). Those situations include urban canyons (in an urban area with dense tall buildings, and GPS signals from satellites used for locating the position would bounce off among the surrounding buildings), tunnels, a heavy tree canopy, and in certain types of vehicles, among others. These situations lead to considerable location calculation delays. Urban canyons may result in significant location errors. Due to the blockage of tall buildings, GPS receivers may not be able to interact with four satellites, producing loss of

[3] A large portion of the users are observed only with a few days and there exist large gaps between consecutive records in a single day.

signal. When a traveler or vehicle equipped with GPS devices takes the subway or goes through a tunnel, GPS devices may not receive any signal from satellites. In addition, according to Gong et al. (2012), warm start (the process when GPS loggers start calculating the current location after being inactive underground) would greatly influence the application of GPS devices for subway and commuter rail mode detection. When a GPS logger is brought from underground to aboveground, it often takes several minutes to relocate the position, during which the traveler may move far away from the subway station, resulting in inaccurate position data. To fix such problems when using GPS devices in urban areas, several technical methods have been proposed in the past, including increasing the number of visible satellites (Cui & Ge, 2003) and augmenting GPS with other navigation systems such as the Russian Global Navigation Satellite System (GLONASS) (Walsh et al. 1997). In addition, incorporating GPS with other positioning methods such as mobile phones or dead reckoning sensors is widely practiced. With the help of cellular networks or Wi-Fi, more complete records can be obtained from GPS devices. Furthermore, most GPS receivers available today have the capability to record inside buildings and obtain position in 5 or less seconds, which also improves the quality of GPS data (Stopher & Greaves, 2007). In addition to those technical improvements, specific algorithms and data-processing methods were proposed, such as finding a constrained solution (Phatak et al. 1999), Kalman-filter-based approach (Ray et al. 2001; Jang et al. 2000), fuzzy logic based map-matching algorithm (Syed & Cannon, 2004), and so on. In the future era of connected and autonomous vehicles, vehicle to vehicle (V2V) and vehicle to infrastructure (V2I) communications may also help solve those problems. As the GPS signal loss issues have been widely reported and studied in the past, and the GPS data obtained for this study do not seem to have obvious signal loss problems, those issues will not be further discussed in this project.

Self-selection

Similar to mobile phone data, GPS data may also have inherent self-selection issues, which may come from at least two aspects.

First, for the same reason that users usually self-select into using GPS devices, data generated from those devices might not be representative for the whole traveling population. GPS data have many sources as aforementioned, including dedicated studies based on GPS, fleets monitoring, navigation devices, and mobile apps. Each source of data may vary in unrepresentativeness. For instance, not every vehicle is equipped with a navigation device. Even when it is, if the driver is familiar with the route (e.g., from home to work), (s)he may not use the device at all. Consequently, data from navigation devices would result in bias if not considering the penetration and use ratio of such devices among vehicles/drivers. Furthermore, dedicated fleets produce GPS data with a broad coverage (or high penetration) of certain group or type of vehicles (e.g., taxi or trucks) , but capture few other vehicles directly.

Secondly, specific information regarding the sources of GPS data is quite limited. As discussed in the categories of GPS data, the data offered by third parties have been popularly utilized in transportation due to their abundance and clear data formats. However, those data are often preprocessed (e.g., for privacy protection), thus they might lack certain critical information, such as vehicle type (private cars, trucks, taxis, shared vehicles, and transits) and unique vehicle identification. Since different types of vehicles often show different travel patterns (for example, delivery trucks often have many “stops” in one day for delivering packages while private cars often

produce regular home- and work-based trips on weekdays), the lack of understanding for vehicle type components and their associated compositions in the fleet would result in bias when processing and using the GPS data.

Lack of Sufficient Details

Similar to mobile phone data, GPS data are often rich in showing "what has happened", but poor in showing "why that happened" (Ban et al., 2014). That is, the data often lack details on behavioral explanations. This may pose challenges when studying and understanding travel related behaviors or inferring mobility related performances by using GPS data only. In this case, supplementing GPS data with other data sources (such as survey data) may be helpful.

3.2.3 Other Issues

Other important, and more general, issues exist related to transportation big data, including mobile phone data and GPS data, such as data collection/sharing, data privacy/security, among others. These issues are not covered in this report.

Table 4. Types of applications and summary of methods and identified issues.

Application types	Methods	Reported issues	Solutions to issues	Examples
Traffic state monitoring				
Traffic speed, travel time	Algorithmic methods to detect trips (matching phone traces with road network)	-Detecting traffic trends over time well, but more accurate when there is no congestion (high speed) -But low accuracy during the AM and PM peak hours	N/A	Smith et al., 2001; Yim and Cayford 2001;Maerivoet & Logghe, 2007; Liu et al., 2008; Bar-Gera, 2007
Urban space monitoring				
Activity locations	Frequency methods (mainly using frequency of visits in each activity location to determine activity locations)	-Predetermined threshold (% of visits) are used to identify activity locations	N/A	Alexander et al., 2015; Phithakkitnukoon, Horanont, Di Lorenzo, Shibasaki, & Ratti, 2010; Wang et al., 2012
	Clustering methods	-Uncertain distance parameter are used as input for clustering methods	Sensitivity analysis, simulation	Calabrese et al., 2010; Ester et al., 1996; Ye et al., 2009
	Statistical model + Algorithm based methods	-Overestimation of activity locations due to the structure of underlying cell tower distribution	Using POI (point of interest) from web can improve the accuracy of activity location inference	Fraley & Raftery, 2002;Chen et al. 2014
Daytime dynamics of population	Erlang (One Erlang is the equivalent of one caller talking for one hour on one telephone)	-Identifying high and low call density area in Milan metropolitan area -Qualitative understandings of day time density of mobile phone usage using visualization methods	N/A	Ratti et al., 2006; Calabrese et al. 2011
Mobility pattern monitoring				
Travel routes	Polygon-based map matching (identifying phone traces' zones and matching it with road network)	-Inaccurate in areas with a dense road network (more than one possible routes) -Sensitive to the types of polygons (geometry) used to search routes	N/A	Schlaich et al., 2010; Ma et al., 2012; Tettamanti et al., 2012 Tettamanti and Varga, 2014; Yue et al., 2014 ; Iqbal et al., 2014; Leontiadis et al. 2014; Bayir et al. 2011; Laasonen 2005

Application types	Methods	Reported issues	Solutions to issues	Examples
Mode choice	-Statistical/machine learning models - Hidden Markov Model/ artificial neural networks - Heuristic methods (rule-based classifiers, Bayes classifiers) -Algorithmic methods (route matching algorithm and K-means clustering algorithm)	-Air travel identified well -Car and transit travel distinguished based on the speed and road and transit network	N/A	Welbourne et al. 2005; Anderson and Muller, 2006; Reddy et al., 2010; Choujaa and Dulay, 2009; Reddy et al., 2010; Stenneth et al., 2011; Widhalm et al., 2012; Feng and Timmermans 2013 ; Nitsche et al. 2014; Yang et al. 2015; Nour et al. 2016

Table 5. A summary of OD methods and identified issues.

Applications types	OD Trips detection methods	Upscaling methods	Reported results and issues	Solutions to issues	Authors/year	Data types (study area)
OD matrix (Dynamic /static ODs)	Network based algorithmic methods (matching mobile phone traces to road network)	- Heuristic method (Scaled up with traffic counts)	-Required to use more detailed zoning system -Representativeness issues (ex, low proportion of elderly people) -Heterogeneity of call rates in different location -Different scaling factor partially address biases	Improving classification of scaling factor may yield better results	Friedrich et al. 2010	FPS data (Germany)
					Iqbal et al. 2014	CDR data (Bangladesh)
					Wang et al. 2012	CDR data (Boston and Bay area in USA)
OD matrix (Static OD matrices)	Cell tower/distance based algorithmic methods (Locational Area/ inferred OD distance)	-Statistical methods -scaling factors (conversion coefficients from econometric model) obtained by comparing the sample with MPO data	-Not able to extract trips within Locational Area, problematics in big Locational Area -Assumption of the stationary time or distance is very sensitive - Limited available input data in the afternoon	N/A	Bonnel et al., 2015	CDR+handover (Paris France)
					Larijani et al. 2015	CDR+Sightings (Paris France)
					Lee et al. 2015, 2016	Social media data (California, USA)

Applications types	OD Trips detection methods	Upscaling methods	Reported results and issues	Solutions to issues	Authors/year	Data types (study area)
OD matrix (OD matrices -- trip purposes based)	Algorithmic methods with an activity location estimation step	-Heuristic method -Scaled up with Census population distribution	-High level (large zone) spatial aggregation produce better results -Biased sample problem were dealt with different scaling factors per OD	N/A	Alexander et al. 2015	CDR data (Boston, USA)
					Serdar Çolak et al 2015	CDR data (Rio de Janeiro, Brazil and Boston, USA)
OD matrix (Dynamic ODs)	Fuzzy logic methods (trip imputation based on activity/passing-by locations)	-Heuristic method -Scaled up with Census Transportation Planning Products (CTPP) data	-Short distance trips and occasional activities without mobile phones are excluded -Almost impossible to capture HOV trips	N/A	Ma et al. 2013	Sightings data (Sacramento, CA)
OD matrix	Commercial software (AirSage's Wireless Signal Extraction (WISE™) technology/ StreetLight software)	-Commercial software (StreetLight) -Several expansion factors (Census data and sighting frequency based)	-Useful for external OD trips analysis and non-residents' trips -Almost impossible to capture HOV trips -Transit trips are ignored -Cellular positioning is the most problematic -Home-based trips are similar but NHB are very different due to the short distance trips -Low fraction of short distance trips (<2 miles) in mobile phone data	Improving classification methods, using bigger zone system, Overestimation of visitors' trips (NHB) in mobile phone, Improving detection algorithm of work place	Milone, 2015	Sightings data Washington D.C.
					Chattanooga 2011	Sightings data (Chattanooga county, TX)
					Fussell et al. 2013	Sightings data (Moore County, NC)
					Huntsinger and Donnelly 2014	Sightings data (Raleigh-Durham region, NC)
					Huntsinger and Ward 2015	Sightings data (western North Carolina)
					Bindra 2016	Sightings data (Syracuse)
					RSG, 2015	Sightings data (Idaho)
					Bernardin et al. 2017	Sightings data (Tennessee)

4.0 Mobile Phone Data

This section presents the three categories of properties (see Table 1) for mobile phone data. Again, the three categories of properties are: zeroth-order properties on describing the data; first-order properties, which are those associated with a trip end or a single activity location; and second-order properties that are associated with trips, or two linked locations or trip ends. Higher-order properties exist for those that involve more than two locations, such as tours, but they are outside the scope of this study.

4.1 Zeroth-order Properties

4.1.1 General Description

The mobile phone data used in this study consists of sightings provided by a commercial vendor. It includes 933,508 users in the Buffalo metropolitan area during the month of April 2014. Sightings are recorded passively for billing or operation purposes when devices connect to the cellular network for calling, texting, internet-accessing and other signaling activities. Each sighting reveals the trace of a user, containing an encrypted mobile identification number, a time stamp and a triangulated location estimation (Table 6). In the report, **a trajectory** refers to the sequence of traces of one user in a one-day period.

Table 6. A sample of mobile phone data.

ID	Time*	Location Record
36cc5**77ca	1396381131	42.95155 | -78.66560
36cc5**77ca	1396385839	42.92989 | -78.62448
36cc5**77ca	1396386054	42.92862 | -78.60799
36cc5**77ca	1396386319	42.92989 | -78.62443
36cc5**77ca	1396386567	42.95155 | -78.66560
36cc5**77ca	1396386653	42.92989 | -78.62443
36cc5**77ca	1396386876	42.95562 | -78.60659

*In Unix time - defined as the number of seconds that have elapsed since 00:00:00 Coordinated Universal Time, 1/1/1970.

Figure 1 gives the density distribution of median of time intervals between two consecutive sightings belonging to the same trajectory (in log-log scale). The distribution decays quickly. About 60% of trajectories have a median of less than 20 minutes. This suggests that sightings recorded tend to cluster together in time, probably around a single event that requires network connection (e.g., phone call, web browsing). The distributions show no clear difference on weekdays and weekends.

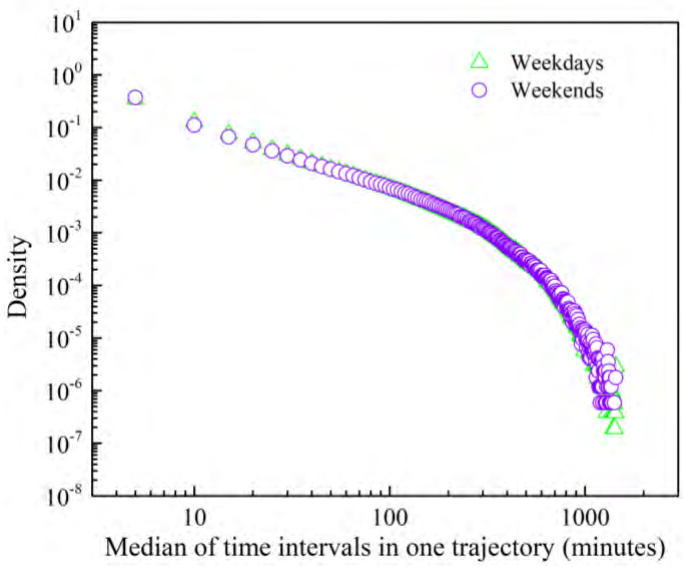


Figure 1. Graph. Density distribution of intervals between consecutive intervals.

The daily distribution of number of sightings shows a weekly pattern (Figure 2), suggesting a daily variation on mobile phone usage. In general, fewer sightings are generated on weekends, indicating less phone usage on weekends than on weekdays.

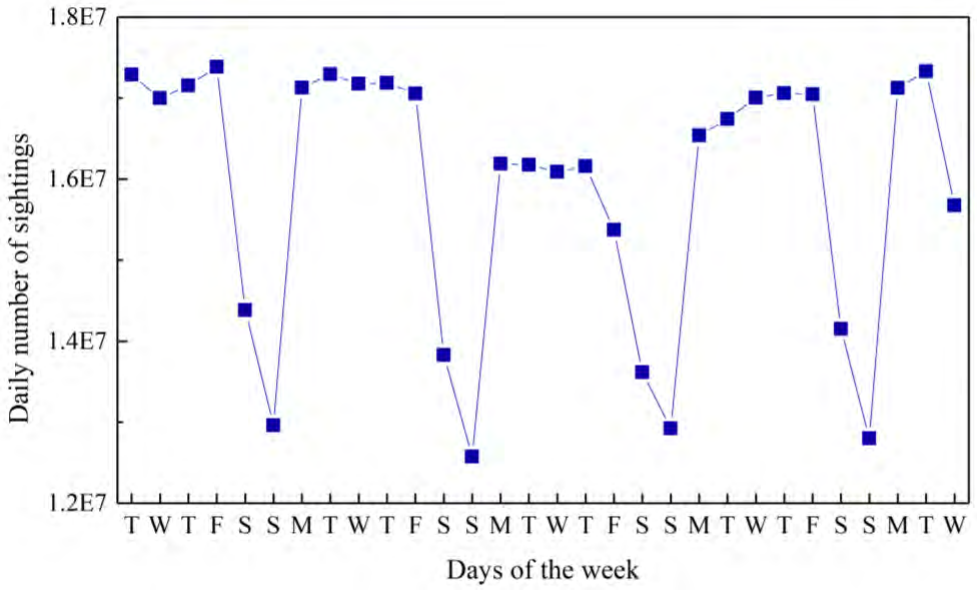


Figure 2. Graph. Weekly pattern of number of sightings.

Figure 3 and Figure 4 show the number of unique IDs in a day in the study period in absolute numbers and fractions. More users are found during weekdays than weekends in general. Fridays tend to have the largest number of unique IDs, whereas Sundays tend to have the least, a finding corroborating with that of Figure 2. On each day, users are categorized by the number of days observed during the study period. Users observed fewer than 10 days account for about 30% in early and late April but their shares are low in mid-April. On the other hand, the opposite pattern

is found for users who are observed between 13 days and 22 days. They account for more than 60% in mid-April, but are less than 40% in early and late April. The users having more than 25 days of records in a month account for 15-25% of all users over a month consistently.

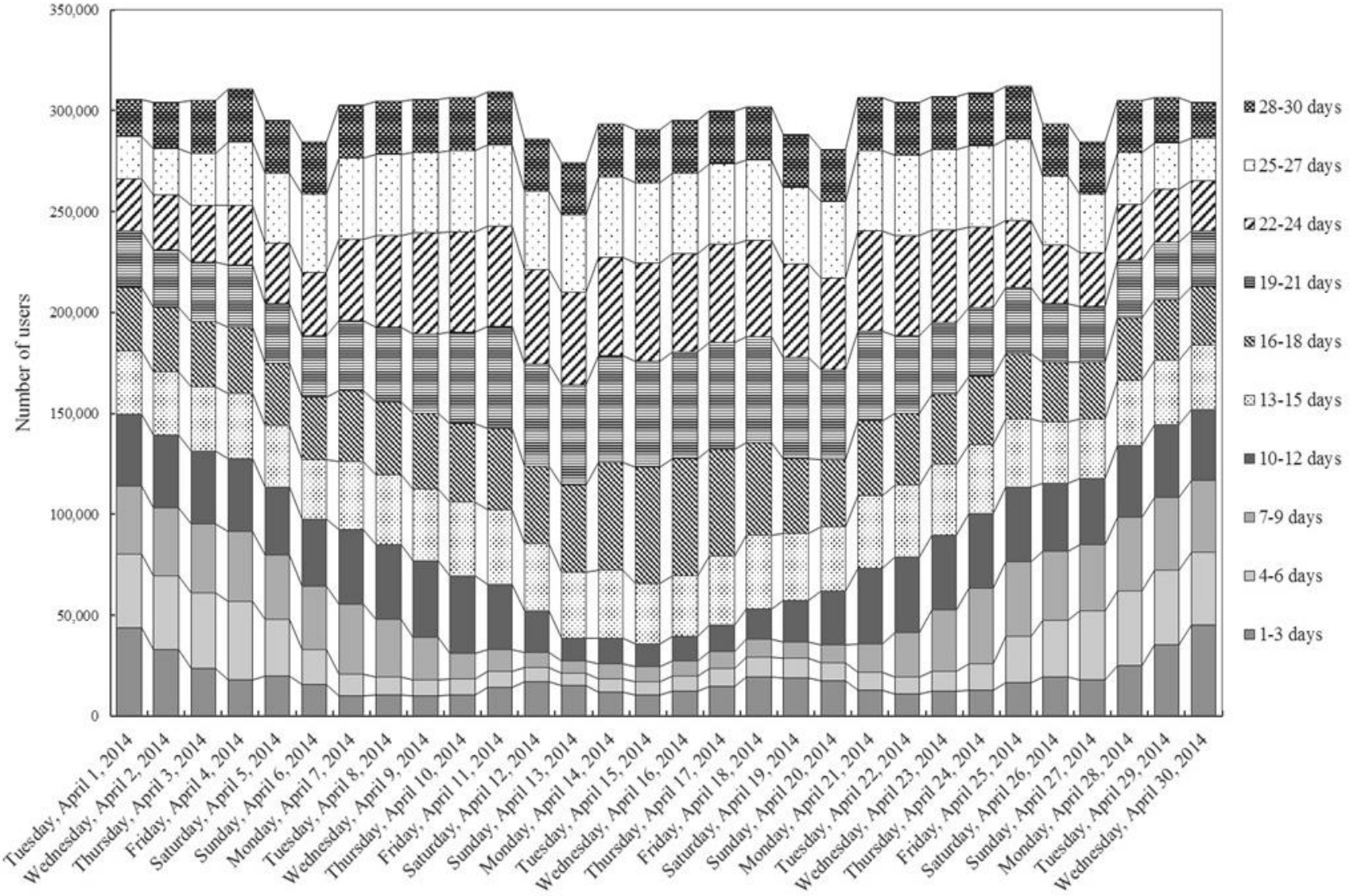


Figure 3. Graph. Number of unique IDs in a day (in absolute numbers).

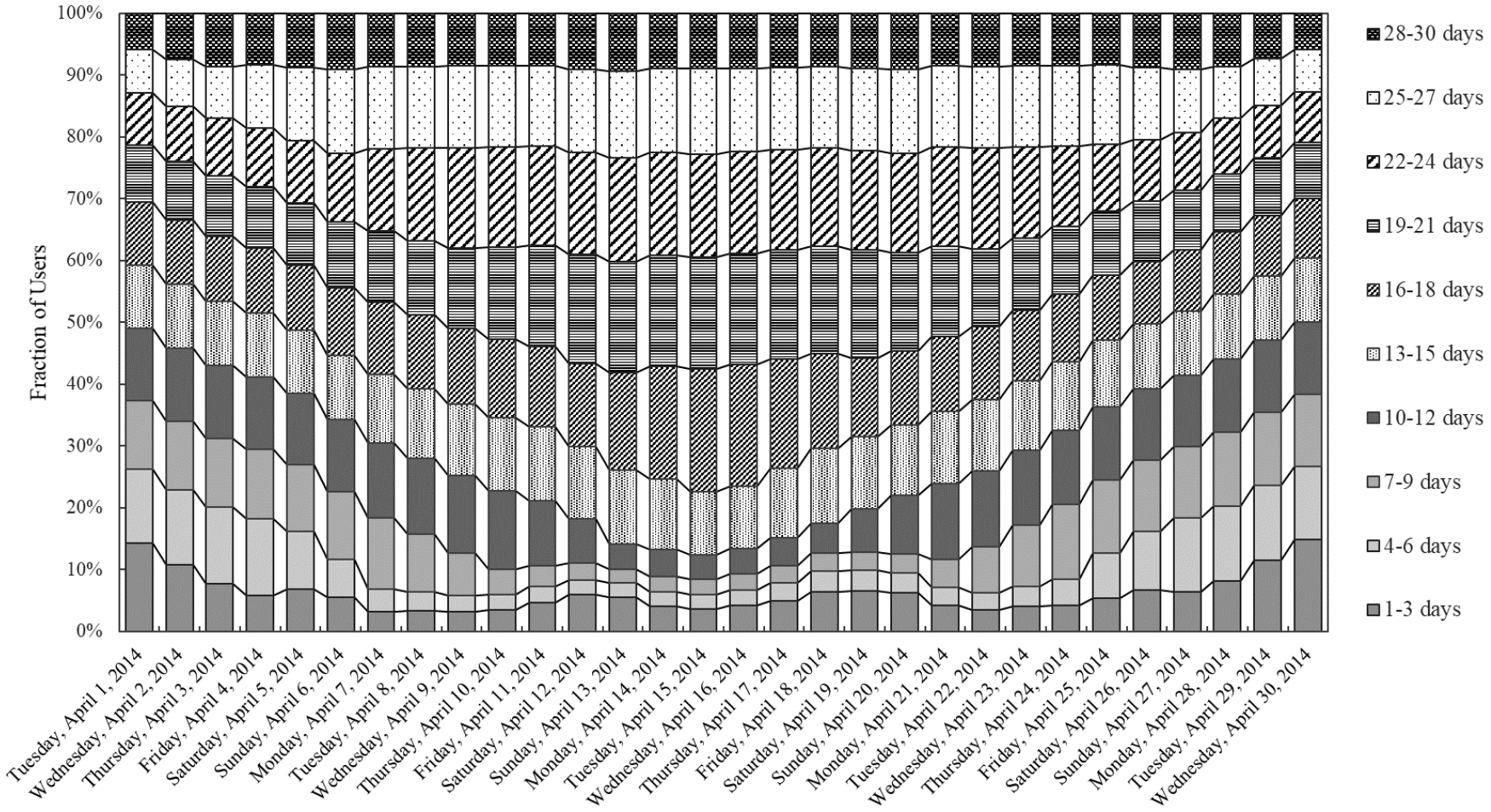


Figure 4. Graph. Fraction of unique IDs in a day (in fractions).

4.1.2 Temporal Sparsity—Interday Sparsity

Previous studies have found that mobile phone data is temporally sparse (Calabrese et al. 2013; Chen et al. 2016; Zhao et al. 2016). In this report, temporal sparsity is investigated with two measures: 1) interday sparsity: the number of days during which users are observed during the study period; 2) intraday sparsity: distribution of sightings throughout a day.

Figure 5 shows the distribution of the number of days observed for all users during the study period. Half of the users are observed fewer than seven days, and only 6,797 users (less than 1% of all users) have at least one trace every day in the entire study period.

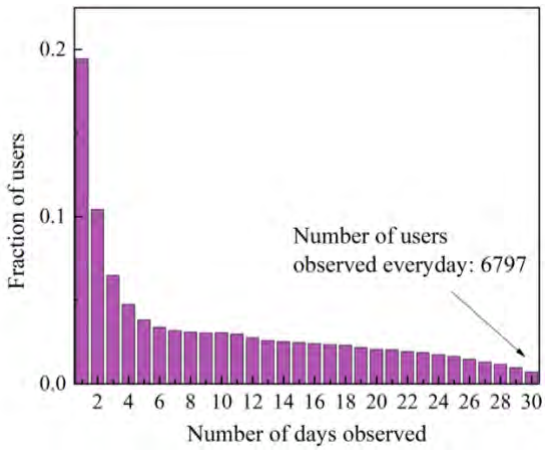


Figure 5. Graph. Distribution of the number of days observed.

Figure 6 shows the distribution of **the life span of unique IDs**. The life span of a unique ID is defined as the time difference between the first day and the last day that a unique ID was observed. Disparity between life span and the number of observed days can be observed. For example, while about 10% of users are observed for two days, only about 6% of them have sightings with a two-days life span. This is because many users observed for two days have a life span longer than two days (e.g., a user may have sightings on April 1st and 3rd, but no sighting on April 2nd). Overall, 63% of the IDs have a life span that is equal to their number of observed days, with the rest having a longer life span.

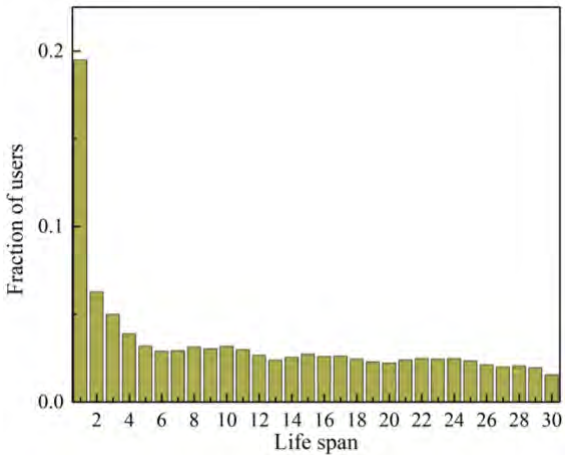


Figure 6. Graph. Distribution of life span of unique IDs.

4.1.3 Temporal Sparsity—Intraday Sparsity

For days with sightings observed, Figure 7 shows how sightings distribute over different times of a day. Most sightings are observed during the midday and early evening on both weekdays and weekends.

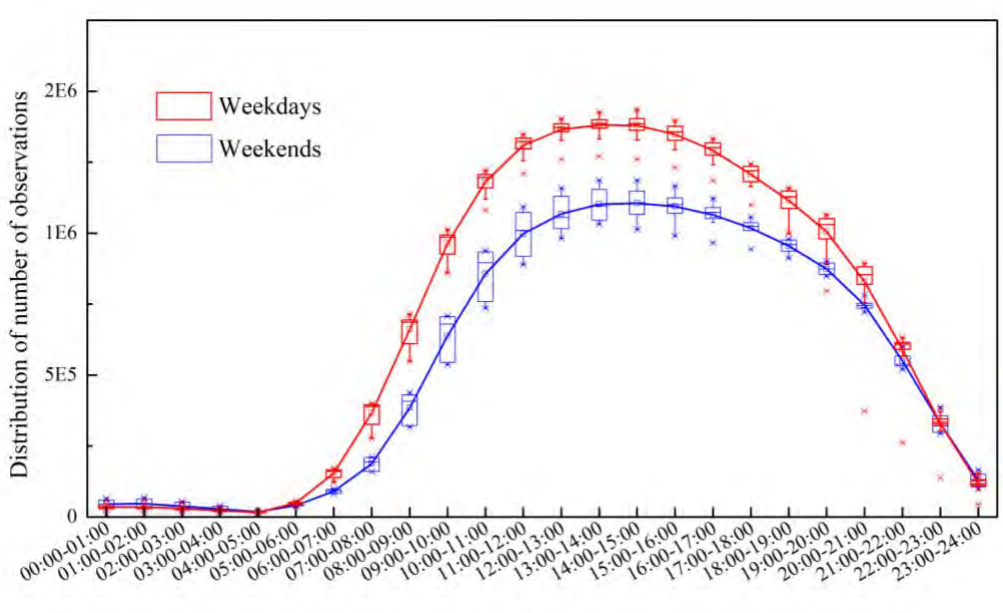


Figure 7. Graph. Distribution of sightings across a day.

4.1.4 Quantifying the Temporal Resolution of a Trajectory

Since sightings tend to cluster in time, a large amount of sightings may not necessarily mean that locations of users are timely recorded. For example, a series of 24 sightings generated during a user's one-hour stay at one location provides less information on the user's movement than someone with sightings evenly distributed throughout a day.

To capture this difference, a day is divided into 24 slots (one hour for each slot). Then **the temporal resolution φ of a trajectory** is defined as, among the 24 hourly slots of a trajectory, the number of slots in which a device is sighted at least once. Figure 8 shows its distribution for all trajectories. The median is six intervals, indicating that half of the trajectories have no more than six hourly slots with their locations revealed.

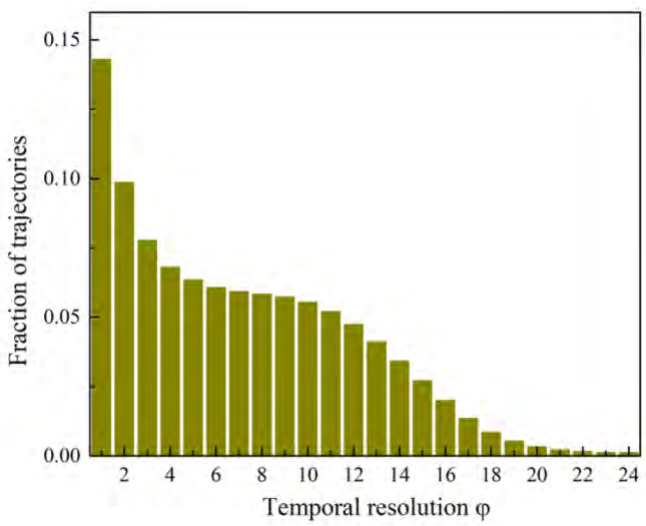


Figure 8. Graph. Distribution of temporal resolution of trajectories.

How these sighted slots (with locations revealed) distribute across various times of a day is further explored. Given a time period of the day, one calculates the fraction of trajectories with their locations revealed at least once. Figure 9 shows how it evolves with time of a day. The evolution curves, both on weekdays and weekends, indicate that it is more likely to observe users' traces during the period from 11:00 a.m. to 6:00 p.m. This single, afternoon-peak distribution suggests that anonymous mobile phone data, such as the one illustrated here, is unlikely to capture peaks in the morning, if it were to be applied to analyze traffic patterns.

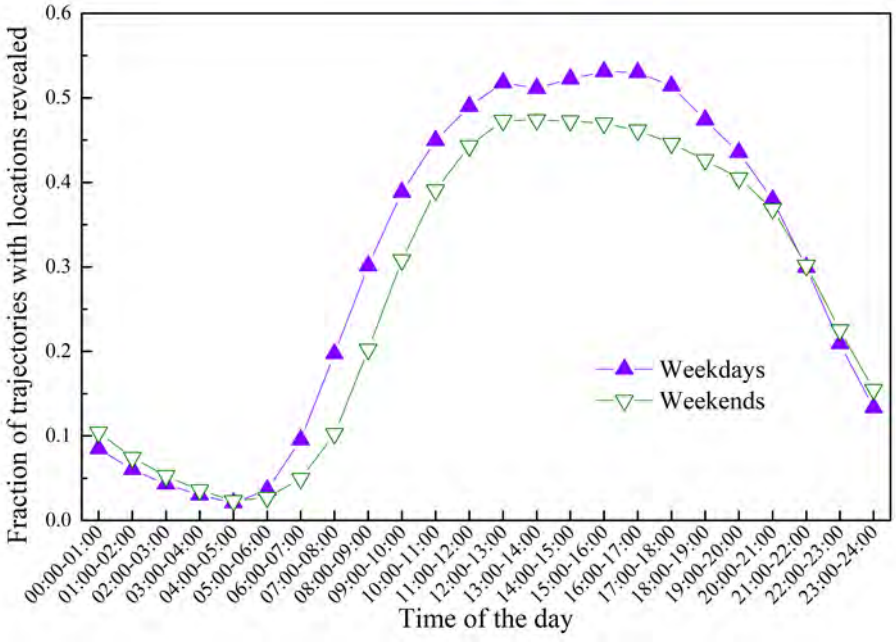


Figure 9. Graph. Fraction of trajectories with their locations revealed at a time of the day.

The temporal resolution of trajectories φ also shows a weekly pattern. Figure 10 gives the average temporal resolution per trajectory on different days of a week, which shows a clear weekly pattern. The temporal resolution remains relatively stable during weekdays, suggesting the regularity of mobile phone usage. However, it drops substantially during weekends, especially on Sundays. This indicates a decrease of mobile phone usage during weekends.

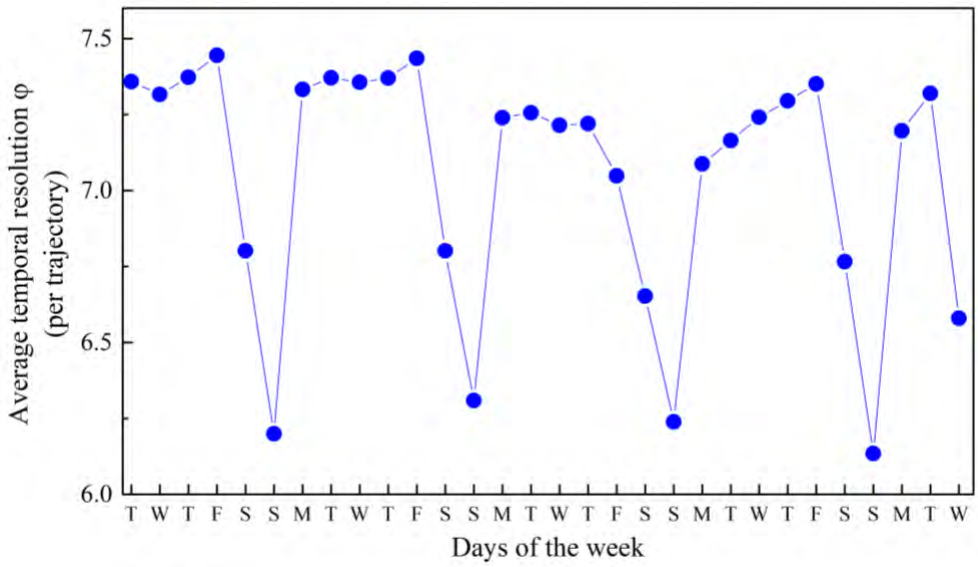
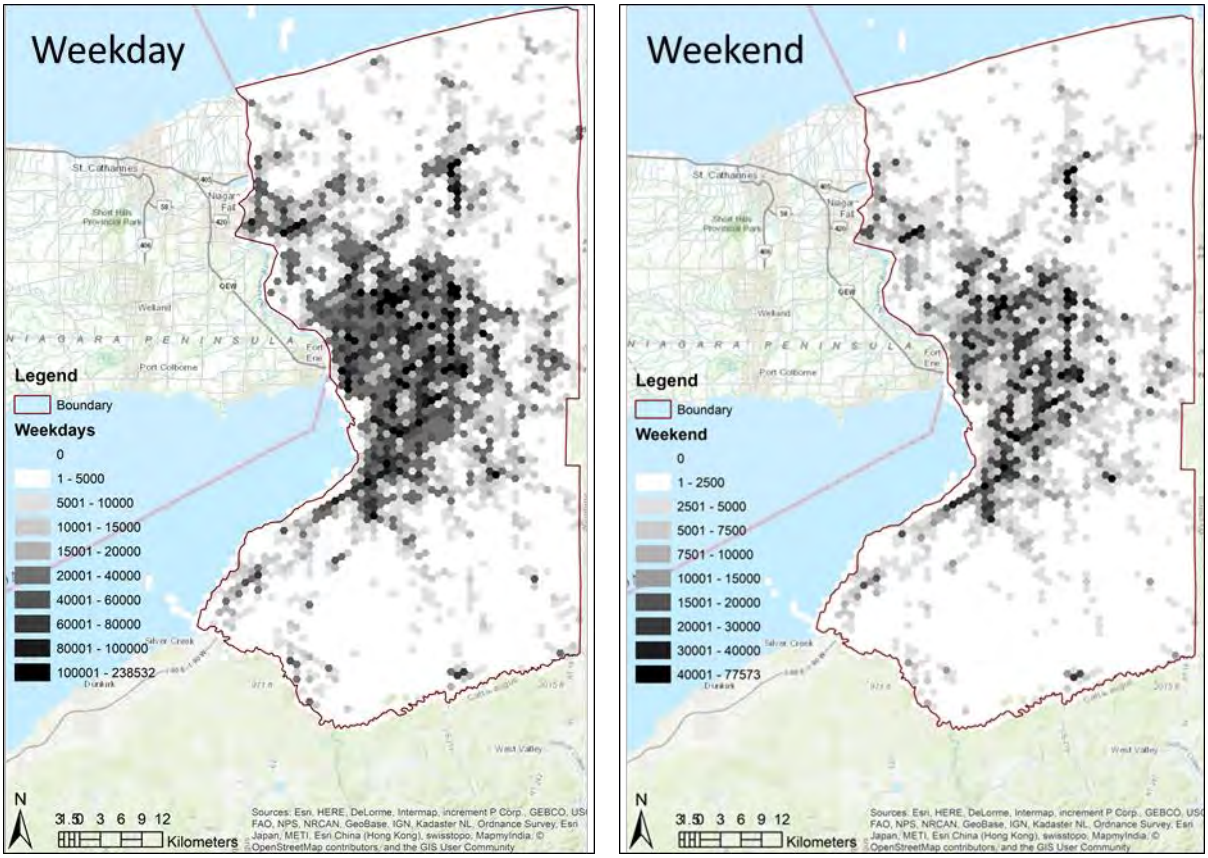


Figure 10. Graph. Weekly pattern of temporal resolution of trajectories.

4.1.5 Spatial Distribution of the Sightings

Both weekday and weekend spatial distributions of mobile phone sightings are shown in Figure 11. Higher concentrations of sightings are depicted with darker color. More sightings are observed and their geographical coverage is larger during weekdays than weekends. Downtown Buffalo is located in the dark-colored center area in the map. During weekdays, many sightings are found both in downtown Buffalo and its adjacent areas and most grid cells in the area are dark-colored. On weekends, the area with dark colors (i.e., more sightings) is smaller indicating less activity as compared to weekdays.



A. Spatial distributions of sightings on weekdays. B. Spatial distributions of sightings on weekends.

Figure 11. Graph. Comparison of spatial distributions of sightings on weekdays and weekends.

Source: © ESRI, World Topographic Map[4]

4.2 First-order Properties

4.2.1 Accuracy of Location Estimation

It was reported that sightings data has an average locational accuracy about 300 meters in urban area. For the mobile phone data used for this study, a plot of certainty radius is created (Figure 12). Certainty radius is a variable representing a radius of a circular area (in meters) where the actual location of a device can be found with a 90% probability. The center of the area is the location record of each trace in the data in the format of latitude and longitude. The statistical mean of certainty radius is 450 meters, with the first, second, and third quartile at 100 meters, 250 meters, and 570 meters, respectively. Figure 12 shows the spatial distribution of certainty radius. Areas with smaller certainty radii are shown in lighter color, indicating relatively higher location accuracy. The downtown area generally shows higher accuracy than the rest of the area due to a higher concentration of cell phone towers.

[4] All maps in this report are created using ArcMap (a product of Esri), unless otherwise specified.

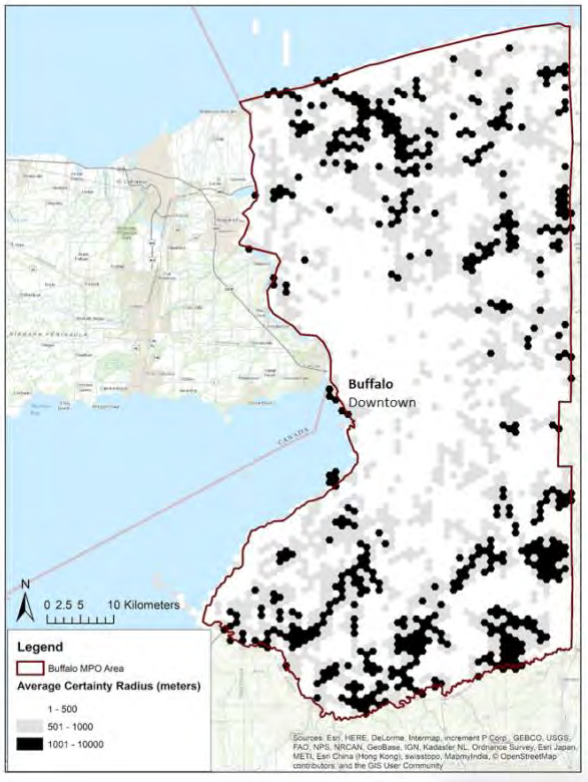


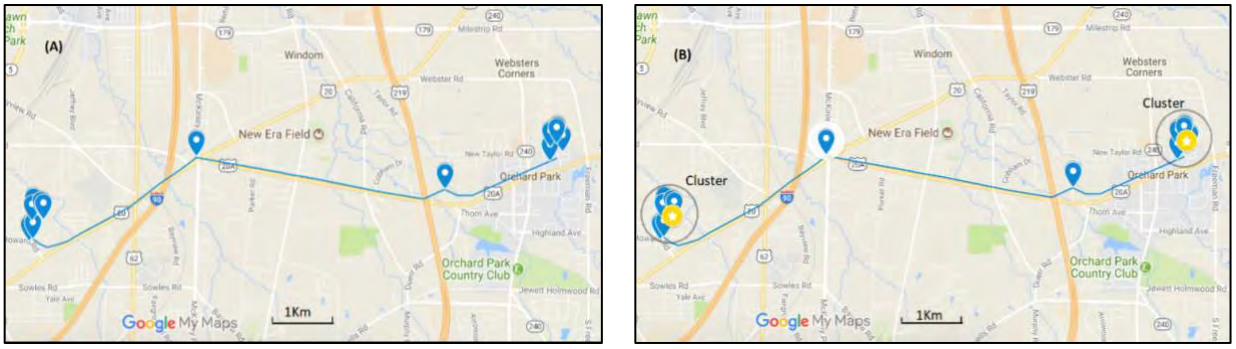
Figure 12. Graph. Spatial distribution of certainty radius.

Source: © ESRI, World Topographic Map

4.2.2 Locational Uncertainty

To derive activity locations from the sightings data, one has to address the issue of locational uncertainty, which originates from the process of location estimation for the sightings data—the triangulation algorithm. Measured values of factors, based on which the triangulation algorithm works, are subject to fluctuations, leading to distinct estimates for the same location. In other words, even when a user stays at the same location, locations recorded at different times vary. This issue prevents the identification of unique activity locations. In addition, it poses difficulty in estimating the amount of time a user spends at a location.

By visualizing traces of one user, one could find that these distinct estimates of one activity location scatter in close proximity (Figure 13-A). A clustering algorithm aggregating these distinct location records is applied to reveal activity locations. The centroid of a cluster is used to represent the activity location where the user stayed (Figure 13-B). See details in *Appendix A* on the clustering method and how to identify activity locations.



A. Locational uncertainty before clustering. **B. Clustering to identify activity locations.**

Figure 13. Illustration. A trajectory of one user showing locational uncertainty.

Source: © Google Map

After the clustering process, **the radius of outputting clusters R_L** is used to measure the extent of locational uncertainty. Figure 14 shows the distribution of R_L. While the majority of R_L fall within 500 meters (as the case shown in Figure 15), there are a small fraction of R_L could be as large as 1000 meters due to the coarse location resolution.

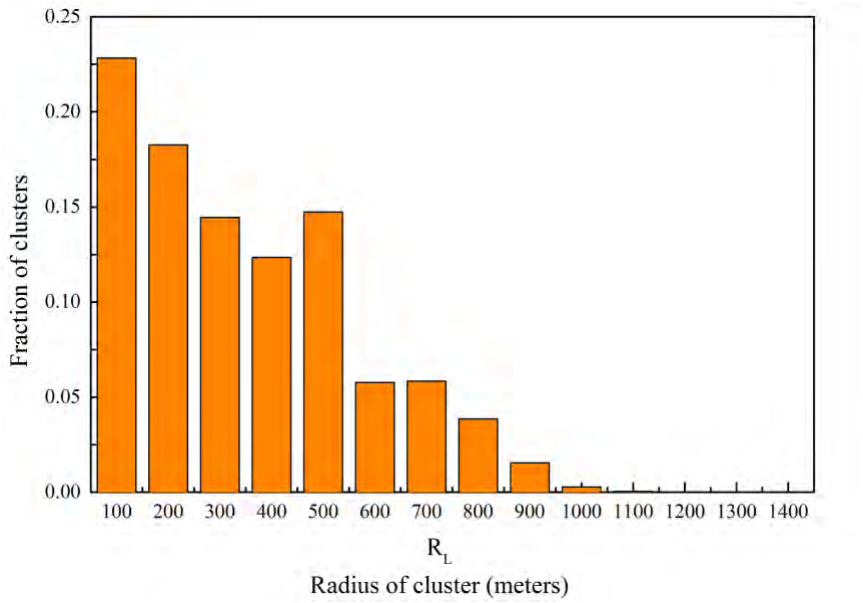


Figure 14. Graph. Distribution of radius of outputting clusters.

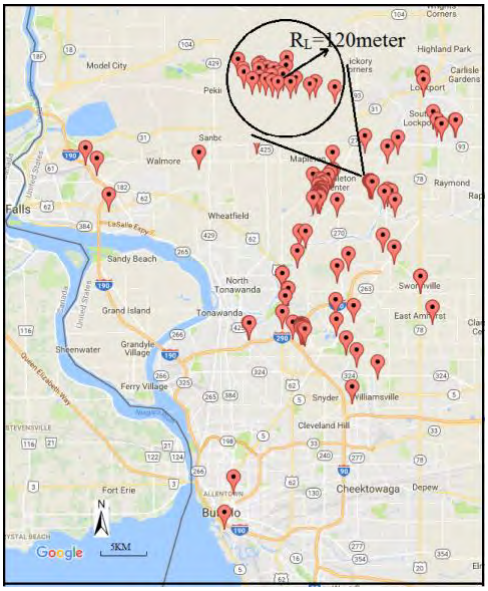


Figure 15. Illustration. A case of locational uncertainty showing that closely distributed location records of one activity location has a cluster radius of 120 meters.

Source: © Google

4.2.3 Oscillation

In addition to the issue of location precision, another issue in the mobile phone data is the oscillation phenomenon (Lee & Hou, 2006; Qi, Qiao, Abdesslem, Ma, & Yang, 2016). It is pure signaling related: some shifts of location records result from signaling activities, as opposed to users' movements. That is, when a user stays at one location, his/her phone may be handed over to different cell towers due to the load-balancing or other operational purposes. When oscillation occurs, some devices may be observed switching between two (or more) faraway locations with high frequencies.

In most cases, an oscillation pattern forms, such as L_0-L_1-L_0-L_1-L_0, where L_0 and L_1 are distinct location records. Table 7 provides an example oscillation case. The Euclidean distance between locations L_0 and L_1 is 2.7 Km, but the user's location switches between them three times within four minutes. If the switching speeds (from one location to another) are calculated, they are incredibly high, which should not reflect actual movements of the user.

Table 7. An oscillation case

Trace	Location	Time	Distance (Km)	Switching Speed (Km/h)
d_0	L_0	12:21:48	\	\
d_1	L_1	12:22:01	2.7	748
d_2	L_0	12:25:20	2.7	49
d_3	L_1	12:25:39	2.7	512

The solution to addressing oscillation phenomenon is necessary as it is such a common phenomenon that traces resulted from oscillation take up a substantial fraction of the total number of records. With a time-window-based method (see Appendix A. Processing Mobile Phone Data),

it is found that the *oscillation ratio* of the mobile phone data is as high as 17% (Figure 16). Here, the oscillation ratio is defined as the ratio between the number of traces resulting from oscillation phenomenon and the total number of traces. The derived results from the data will be biased without a data-processing step to removing oscillation traces.

The oscillation pattern could be complex: an occurrence of oscillation phenomenon may not simply involve only two locations, as the case in Table 7, but involve more than two locations. Figure 16 gives the distribution of different oscillation patterns detected in the data, which are indicated by the number of locations involved in the occurrence of oscillations. In most cases, oscillation only involves two locations; mobile phone data contains complex oscillation patterns: among the 17% oscillation traces, 3% of them are resulting from oscillation involving at least three locations (Figure 16). These complex patterns suggest that pattern-based methods could be less effective (Iovan, Olteanu-Raimond, Couronné, & Smoreda, 2013; Qi et al. 2016; Wu et al. 2014).

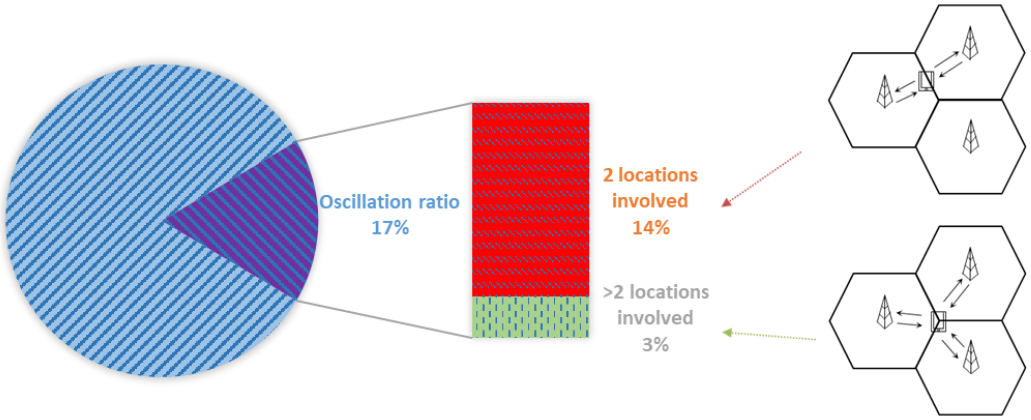
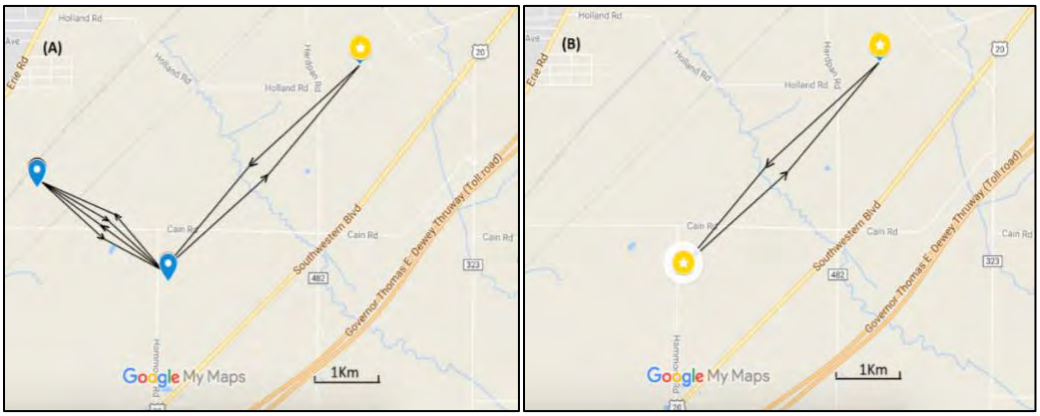


Figure 16. Chart. Oscillation ratio and complex oscillation patterns.

Without removing oscillation traces, observations from the derived trajectories could be biased. For instance, without removing the oscillation traces, three activity locations are derived in one trajectory (Figure 17-A). However, after removing the one generated due to the occurrence of oscillation, two activity locations remain in the trajectory (Figure 17-B).



A. Before removal of oscillation traces. B. After removal of oscillation traces.

Figure 17. Illustration. A trajectory of one user with oscillation traces.

Source: © Google

4.2.4 Activity Duration

A trajectory is typically featured with a sequence of travel and stays at activity locations. Due to the temporal sparsity discussed above, as demonstrated in Figure 18, the observed arrival time $\hat{t}^i_{arr}$ from one activity location i may not the actual arrival time t^i_{arr}. This is also true to the observed departure time $\hat{t}^i_{dep}$. Clearly the observed activity duration, which is defined as the time difference between the observed departure and arrival time, could be a biased estimation on the actual activity duration.

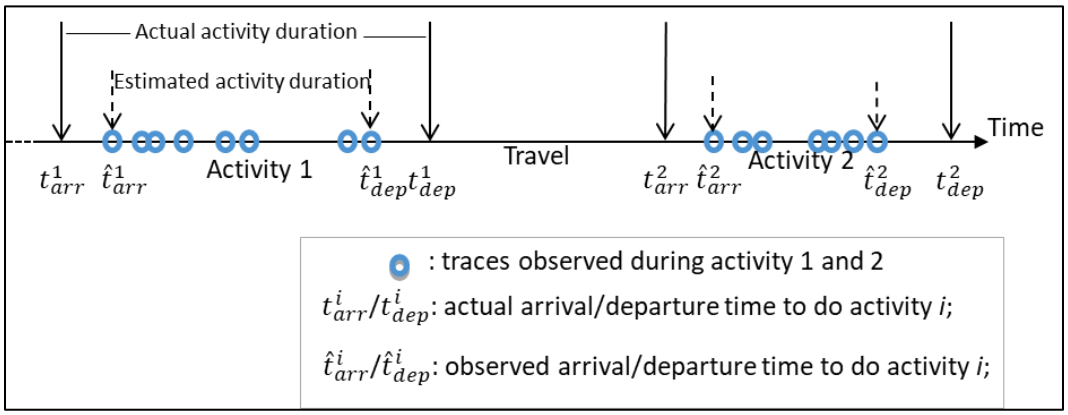


Figure 18. Illustration. Activity duration and a demonstration on the biased estimation.

Figure 19 gives a comparison of activity duration derived from mobile phone data on weekdays and travel survey data (2002 Greater Buffalo-Niagara Transportation Survey, 2003). The comparison offers two observations. First, there are fewer activities of short (less than 1 hour) and long (exceeding 5 hours) duration in the mobile phone data than in the household travel survey; second, mobile phone data has more activities with duration between 1 and 5 hours than the household travel survey. Only those observed stays (with duration longer than 5 minutes) are included in the analysis. And no clear difference is observed on the distribution of activity duration between weekdays and weekends (Figure 20).

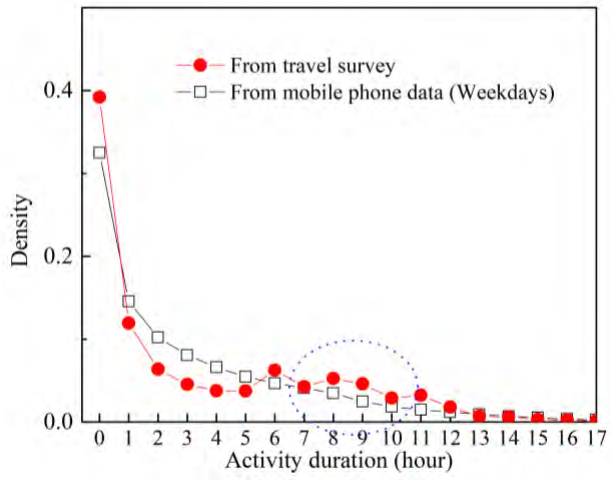


Figure 19. Graph. Comparison of activity duration derived from travel survey and mobile phone data.

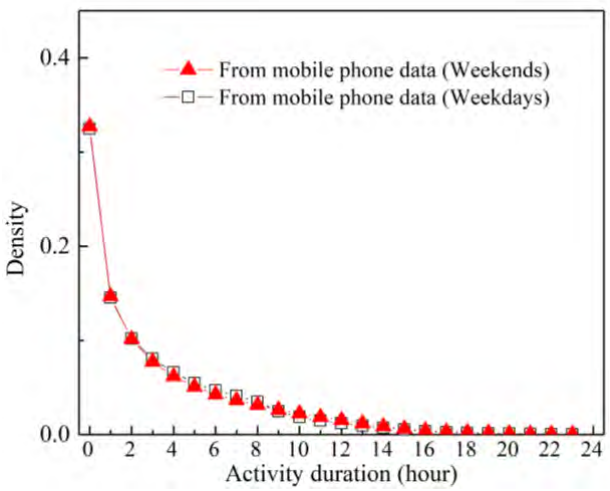


Figure 20. Graph. Activity duration observed on weekdays and on weekends from mobile phone data.

4.2.5 Identifying Home Locations

To identify the resident population from the mobile phone data, home locations of the users in the mobile phone data must first be inferred. For each user, his/her total number of unique activity locations, as well as frequency of visits to each location and the associated activity duration throughout the study period are different (Wang & Chen, 2017). For home locations, sightings observed during night time and the last sighting location of a day are important information as one usually spends the night at home and returns home at the end of a day. In this study, night time is defined between 6 pm to 6 am the next day. For each user, a score is calculated for each location based on three statistics: the number of times a location being visited at night, the amount of time spent at a location, and the number of times a location is being observed as the last location in a day. For each statistic, a rank from 1 to n, where n is the number of unique locations a device is observed throughout the study period, is calculated. Then, the sum of the ranks across all three statistics gives the score. As an example, we consider a user who has 3 activity locations (*a*,*b*,*c*). He/she visited each location at night at the frequencies of 3, 15 and 5 respectively, stayed in each location at a total of 5, 10, and 30 hours during the study period, and was observed being the last location of a day at a frequency of 2, 5, and 1 during the study period. In this case, the scores for the three locations would be 4 (1+1+2) for *a*, 8 (3+2+3) for *b*, and 5 (1+3+1) for *c*, respectively, and thus, the second location *b* is inferred as the home location. In the case where the scores for two locations are the same, the frequency of visits to the location during day time is also used to identify the home location. There are some visitors (presumably) who do not have enough records (for example, no records at night time) to infer home locations. In this case, the most frequently visited location during daytime is selected as home.

4.2.6 Comparing Against Census Data

Figure 21 illustrates the spatial distributions of population density and inferred home density in census tracts. The correlation coefficient between the two spatial distributions is 0.433, and is less than 0.4 when population size and number of inferred homes are compared. In Figure 22, all census tracts are classified based on the median values of population density and the inferred

home density. The high population density and high inferred home density zones and the zones with the low densities are colored with red and dark blue, respectively. It is plausible that more users' homes are found in the zones with high population density. However, the other two types of zones pose considerable concern: the zones with high density of population but low density of inferred home (green) suggest likely under-representation and reversely, zones with low density of population but high density of inferred homes (yellow) suggest likely over-representation. This may be due to the fact that mobile phone data contains a substantial number of nonresidents such as visitors, as evidenced by a large proportion of users with short days of observations (Figure 5) and lifespan (Figure 6).

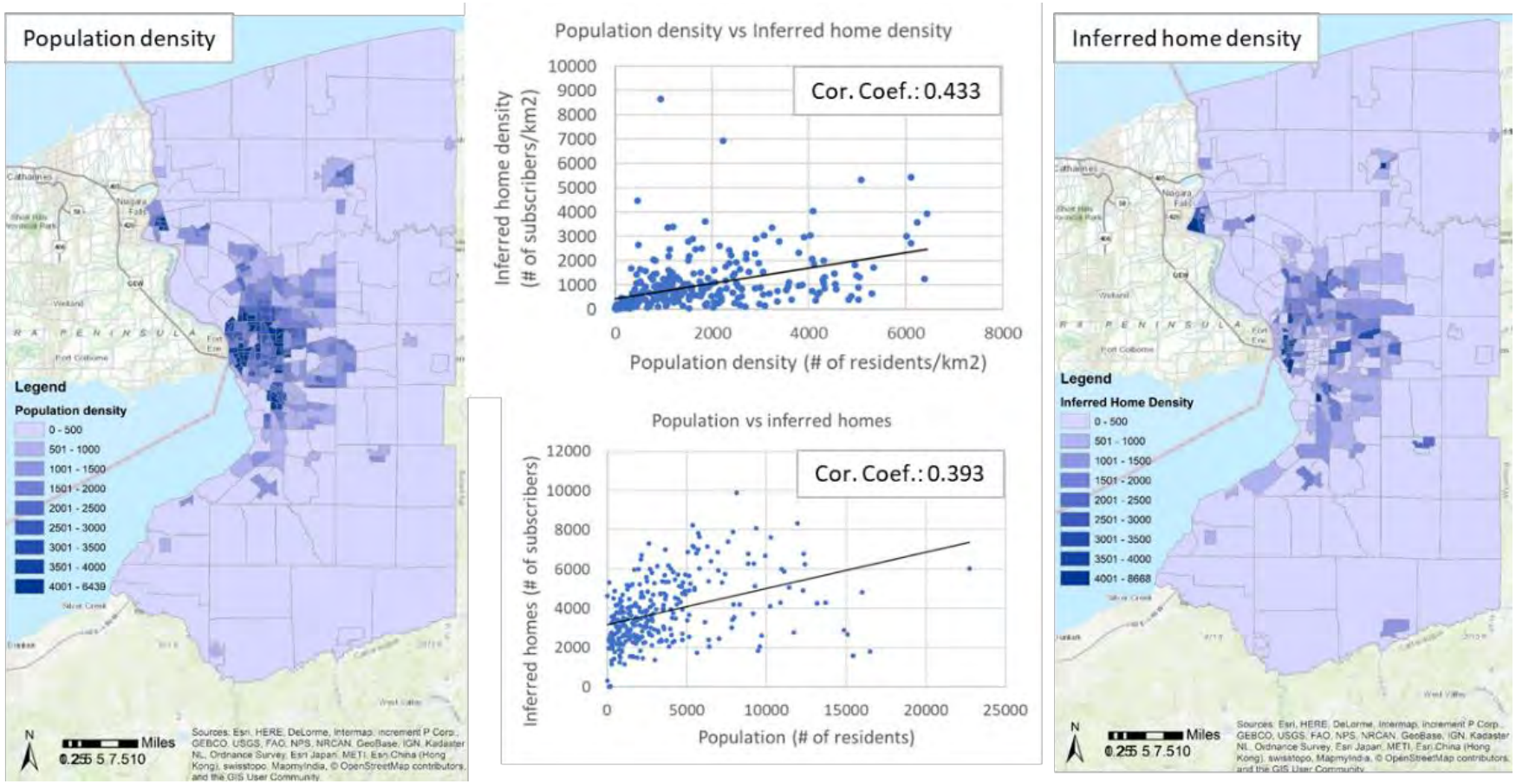


Figure 21. Graph. The spatial distributions of Census population and inferred home locations.

Source: © ESRI, World Topographic Map

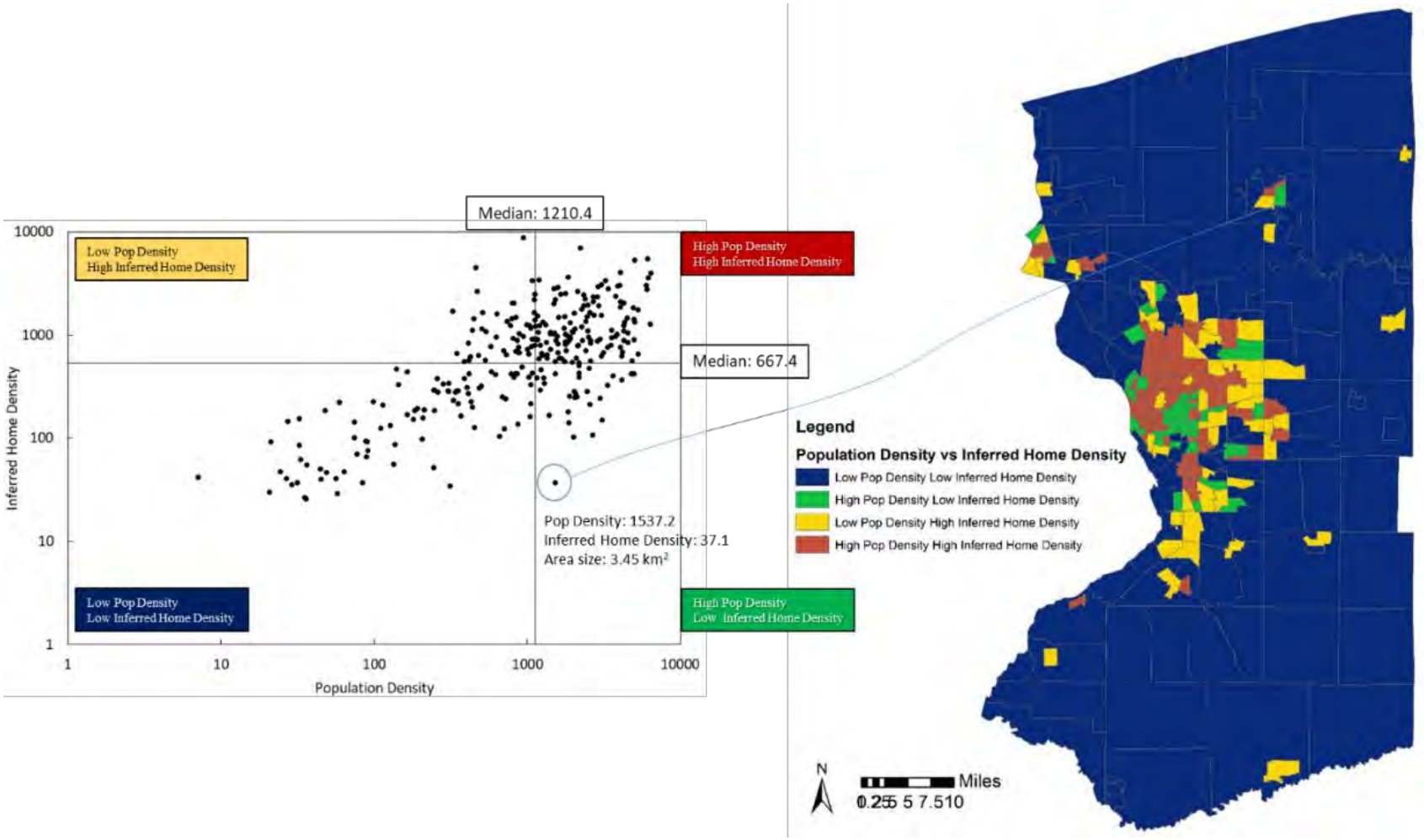
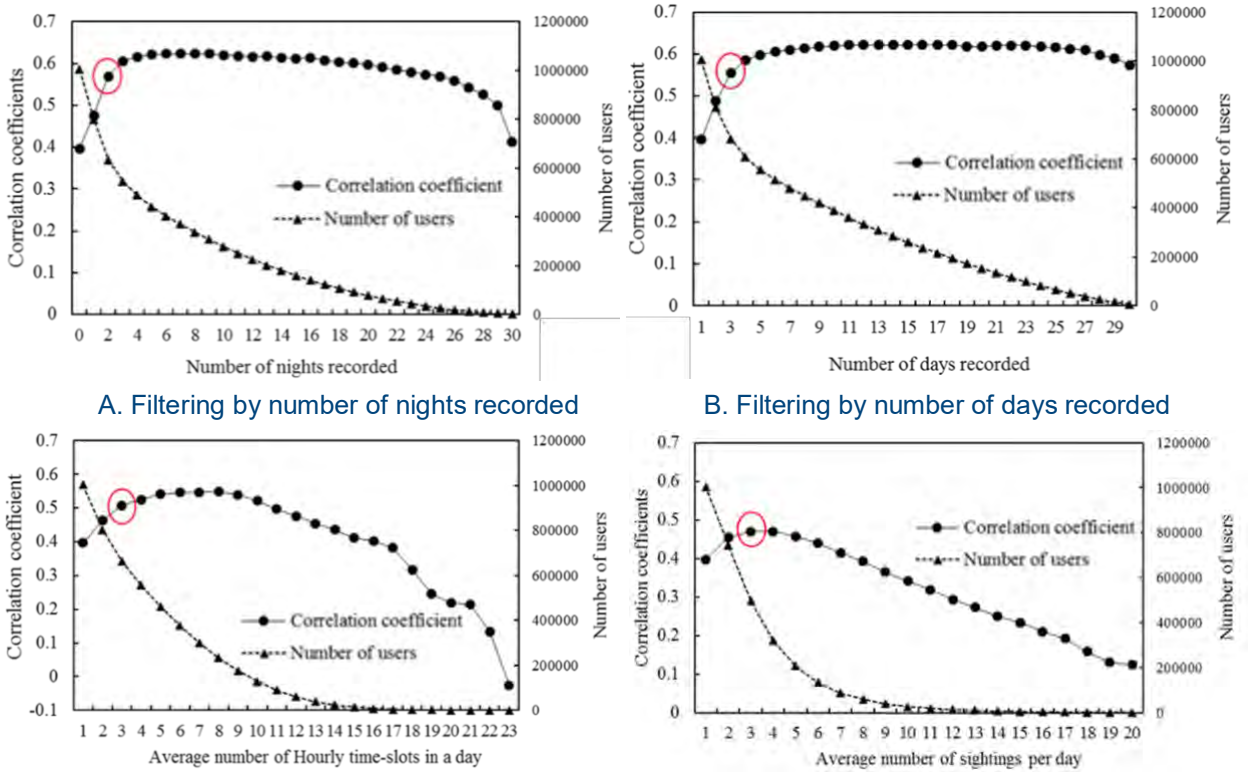


Figure 22. Graph. The comparison between Census population and inferred home locations.

In order to filter visitors out of the mobile phone dataset, four criteria are used including the number of nights in a month, the number of days in a month, average temporal resolutions per user per month[5], and number of sightings in a day. Without using any filtering criterion, the similarity between population size and inferred home distribution is about 0.4 (Pearson correlation coefficients). Figure 23 shows the changes in correlation after using different filters, and the number of remaining users after applying the filters. The filters of the number of nights and days appear to be more effective than the other two since the correlation between the two distributions increases by 0.2 when the filters of more than 3 nights or days are applied. This indicates the visitors staying at hotels can be removed at least to some extent from the mobile phone data sample with these filters. At the same time, however, low temporal resolutions of residents' records can increase the difficulty in inferring their home locations. Thus, although applying the temporal resolution and number of sighting filters do not increase the correlation as much as the other two filters, they are still helpful to identify users with low temporal resolutions and increase the correlation by excluding them. As a result, four criteria are used to infer home locations: more than three nights, more than three days in a month, more than three temporal resolutions in a day on average, and three sightings records. As a result of applying all these criteria together, the correlation coefficient is 0.642 and the remaining number of users is 297,718.



A. Filtering by number of nights recorded

B. Filtering by number of days recorded

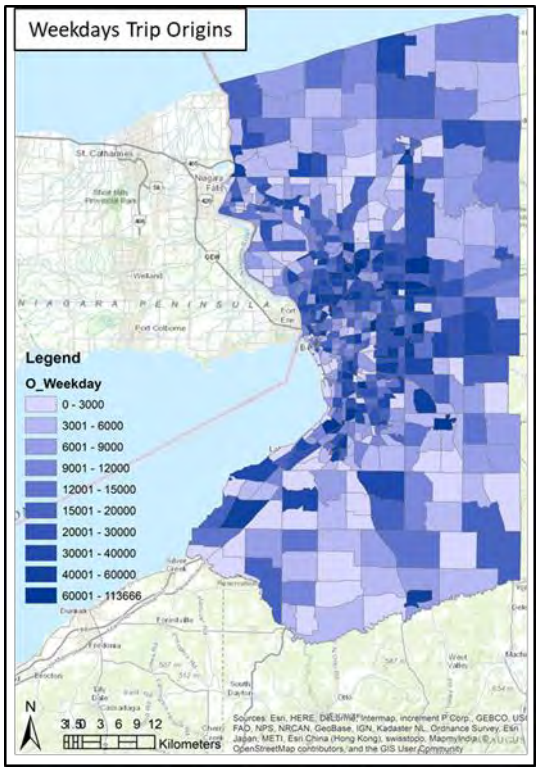
C. Filtering by average temporal resolution

D. Filtering by average number of sightings

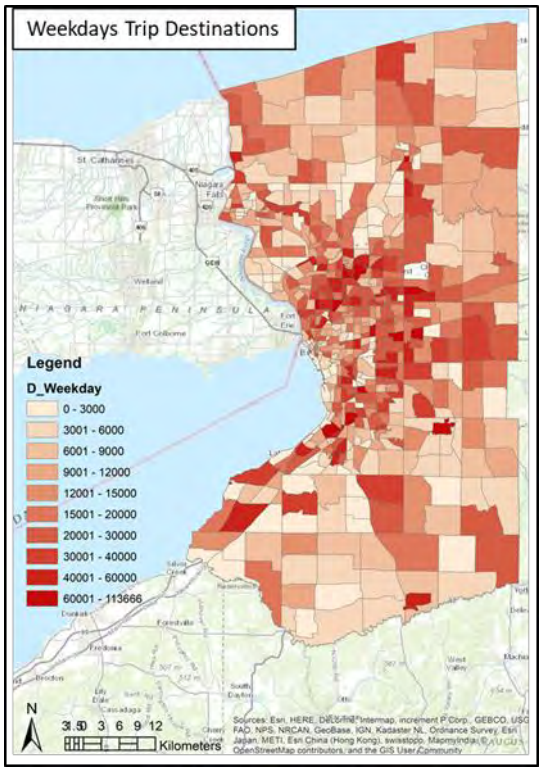
Figure 23. Graph. Applying individual filters to select the best sample from all users.

[5] This is defined in Section 4.1.4. This is average number of hourly time-slots (with at least one sightings) per day per users. Therefore, 1 time-slot means a user is only observed during 1 hourly time-slot in a day.

Figure 24 illustrates spatial distributions of trips origins and destinations on weekdays. In both (origins and destinations), downtown Buffalo has more trip ends—for example, the number of trip origins in downtown Buffalo area is between 15,000 and 113,666 while most of the rest of the area is below 15,000. In general, the distributions of trip origins and destinations on weekdays are similar to each other. However, there is a disparity when they are compared with distributions on weekends (Figure 25). During the weekends, the trip origins and destinations are less concentrated in the downtown Buffalo area and many zones in this area are marked with brighter blue and red colors (mostly less than 3,000). On the other hand, zones in the rest of the study area have more trip origins and destinations.



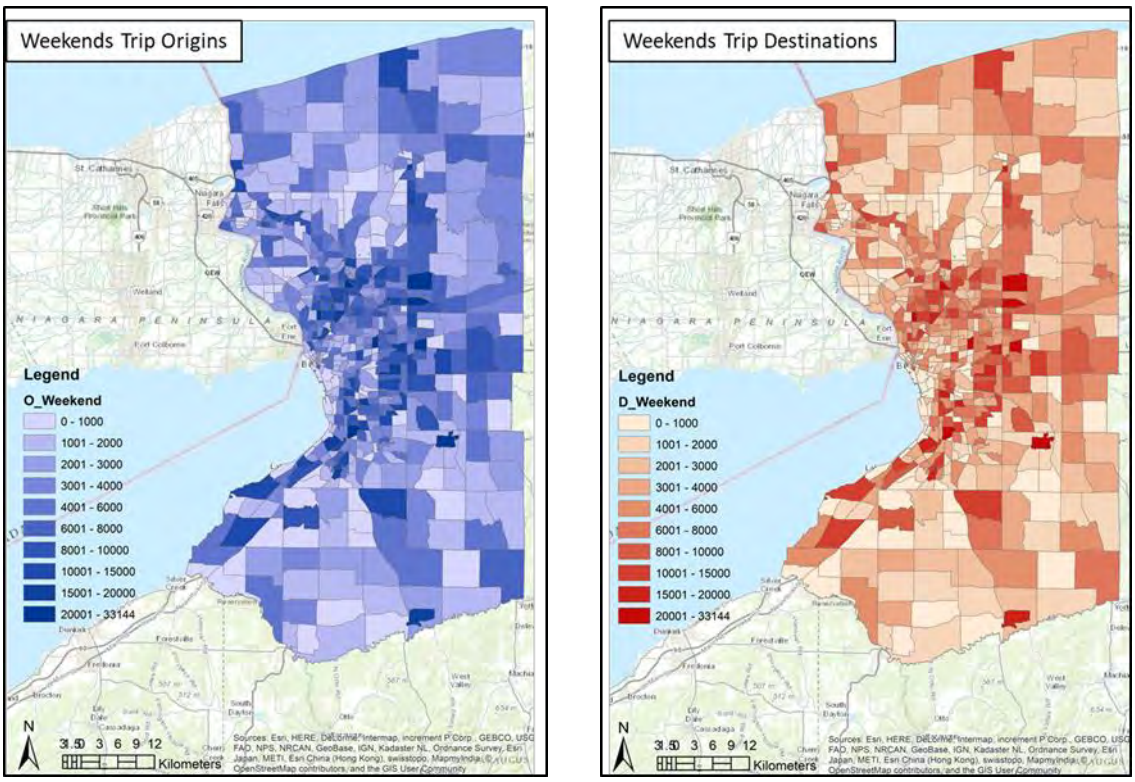
A. Spatial distribution of weekday trip origins.



B. Spatial distribution of weekday trip destinations.

Figure 24. Graph. Spatial distribution of weekday trip origins and destinations.

Source: © ESRI, World Topographic Map



A. Spatial distribution of weekend trip origins. B. Spatial distribution of weekend trip destinations.

Figure 25. Graph. Spatial distribution of weekend trip origins and destinations.

Source: © ESRI, World Topographic Map

To examine the temporal variation of spatial distributions of trip origins and destinations, a correlogram of trip origins and destination in each day of a month is computed. The correlogram shows the similarity between each zone's value (in terms of trip origins and destination) and the mean of its neighbors' values, and the neighbors are defined based on the distance between zones. Therefore, this correlogram illustrates degrees of the spatial autocorrelations of trip origins and destinations depending on the distances among zones. In Figure 26, the correlograms of trip origins derived from mobile phone data are classified into two groups, one for weekdays and the other for weekends. Figure 26 shows that these two groups are well differentiable, with weekday correlograms generally higher than weekend ones. This suggests that the spatial autocorrelation of weekday trip origins is more positive than that of weekend trip origins. These correlograms of trips origins derived from mobile phone data are compared with the one from MPO results (models). The comparison suggests that the spatial autocorrelations of weekday trip origins are more similar to the MPO results than the weekend trips (Figure 26). This may be related to the fact that the MPO results were estimated based on the weekday trips only. The similarity between the zones and their closest neighbors in mobile phone data (indicated by distance scope 0-3km) is much lower than the MPO results. In other words, there are substantial differences between trips from/to the zones and their closest neighbors in mobile phone data and in the MPO results. This is presumably because of the missing or neglected short-distance trips in mobile phone data

collection process or activity location estimation process. As shown by Figure 27, similar patterns are observed on correlograms of trip destinations.

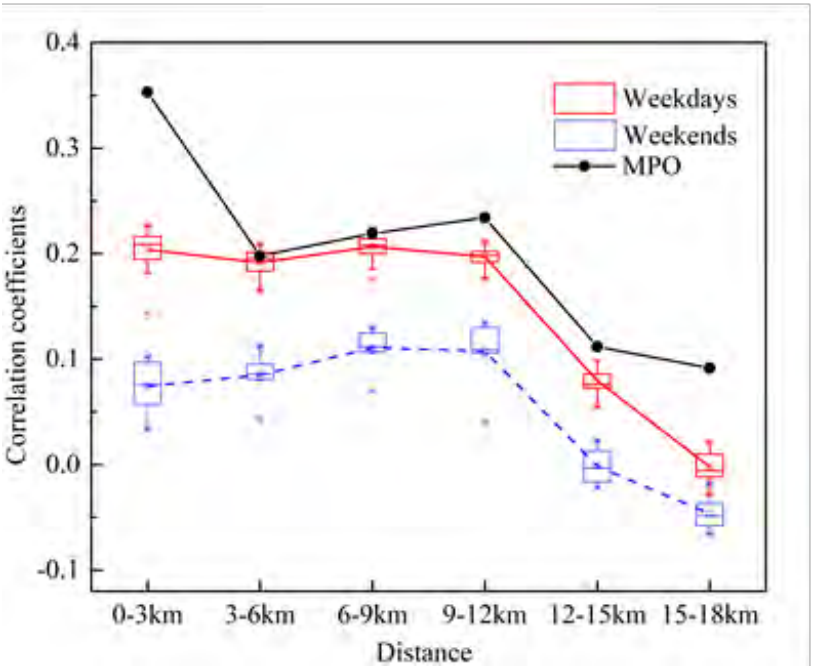


Figure 26. Graph. Spatial autocorrelations (correlogram) of trip origins.

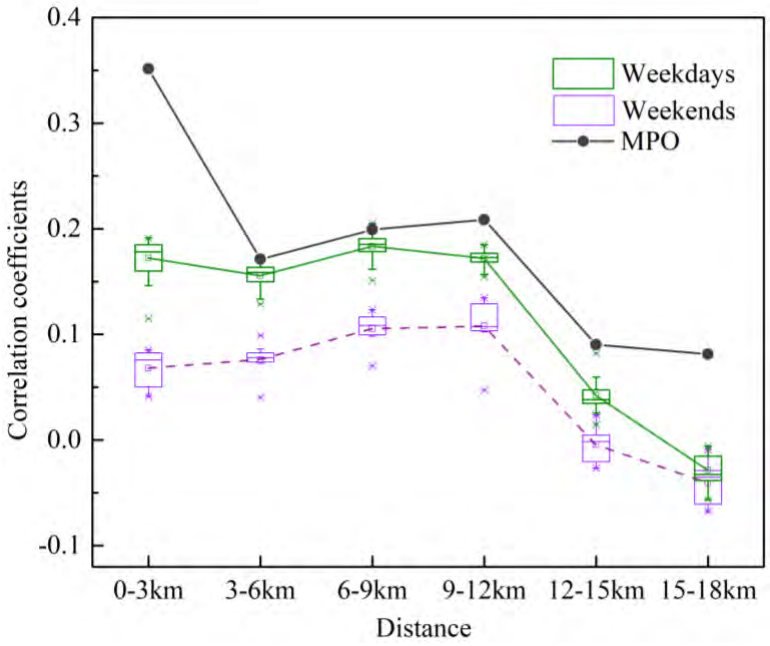


Figure 27. Graph. Spatial autocorrelations (correlogram) of trip destinations.

The spatial distributions of the trips extracted from mobile phone data (MPD) trips are compared with MPO model results and population distributions in (Table 8). The highest correlation

coefficients are found in the relationship between MPD trip origins/destinations and MPO model results during the weekdays (0.690 and 0.682). The relationship between MPD trips and population distribution, however, shows a much lower degree of similarities. This is likely because the trips origins and destinations include not only users' homes but also their other activity locations (0.543 and 0.472). For example, there are several zones with a large number of trip origins and destinations (60,000-120,000 for 22 weekdays), but the number of residents in these zones is under 4,000. During the weekend, the similarities between MPD trips and MPO results are slightly lower than weekday ones. As discussed earlier, this is likely because MPO model results are estimated only with weekday trip records. The similarities between mobile phone trips and population during the weekend are much lower than the ones during the weekdays.

Table 8. The correlations between trip origins and destinations of mobile phone and MPO results /and population counts.

	Corr. Coeff. Mobile phone vs MPO	**Corr. Coeff. Mobile phone vs Census population**
Trip Origin (Weekday)	0.69	0.543
Trip Destination (Weekday)	0.682	0.472
Trip Origin (Weekend)	0.665	0.403
Trip Destination (Weekend)	0.66	0.396

The daily variations of these relationship are shown in Figure 28 and Figure 29. The correlation coefficients between trip origins and destinations in MPD and MPO model results range between 0.680 and 0.700 on weekdays, and are slightly lower during weekends (0.660-0.670) except for April 20, 2014 (the 3rd Sunday). The daily variation of the comparison between population and trip origins and destinations is drawn with dotted lines with triangular markers. Interestingly, the trip origins during weekends are more similar to population distributions than weekdays, but the opposite patterns are found in the relationship between trip destinations and population. The only exception is the correlation coefficient between trip destinations and population on April 20, 2014.

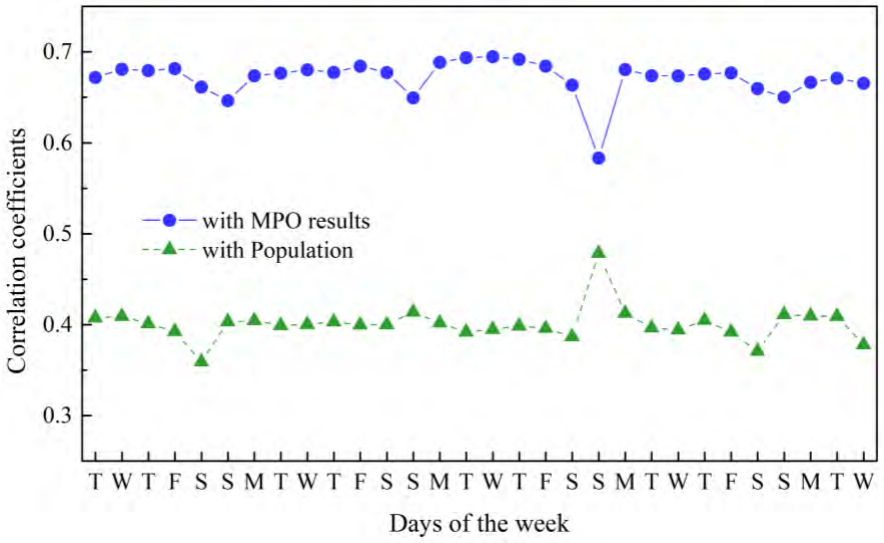


Figure 28. Graph. Daily variations of correlation coefficients between trip origins derived from mobile phone data and MPO results /and population counts.

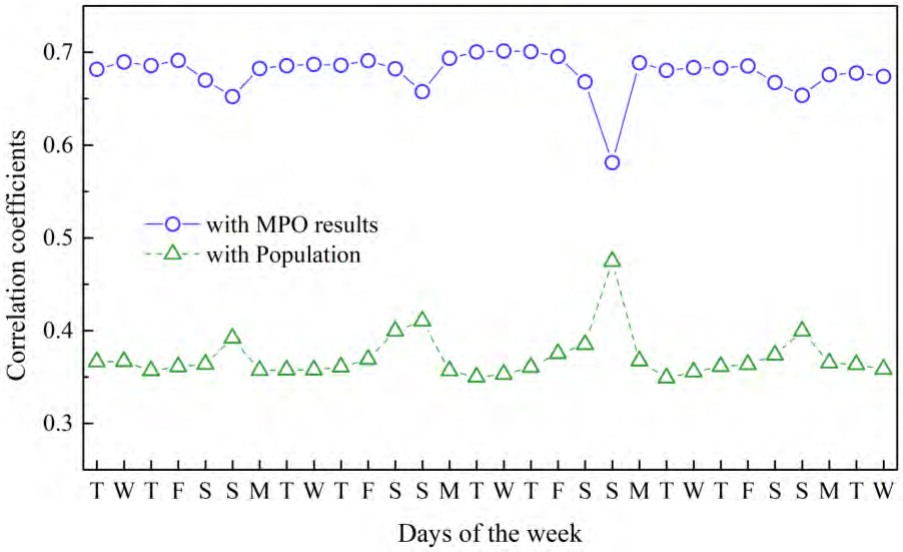


Figure 29. Graph. Daily variations of correlation coefficients between trip destinations derived from mobile phone data and MPO results /and population counts.

4.2.7 Capturing Evolutionary Characteristics

Due to the temporal sparsity, we may fail to observe all stays visited by users. The failure could be compensated by taking advantage of the longitudinal nature of MPD: given the study period of one month, the frequently visited locations ought to be sighted and revealed by the network. This is rooted in the regularity of individuals' travel behavior (e.g., preferable returns) (González et al. 2008; Song et al. 2010; Jiang et al. 2013). The probability of visiting the first three most visited locations could be as high as 0.85 (González et al. 2008). However, with a short observation period, visits to these locations may not be observed due to the temporal sparsity. Therefore, it is expected that the observed probability of visiting anchor locations would evolve as the observation period grows.

To check this, **the observed regularity of visiting anchor locations R(t)** is defined as the fraction of trips ending at the first three most visited anchor locations among all trips observed up to time t. Figure 30 shows the evolution of R(t) with the length of the observation period during which only users observed up to that period are used to calculate R(t). The substantial jump from when using data of one day to when using data of two days suggests that those frequently visited locations are unlikely to be derived from data of a single day. The drops on Fridays and weekends are attributed to users' explorations to new activity locations (i.e., visits to new locations). A weekly pattern is also observed here with a gradual decrease until after three weeks. This indicates that missed visits due to the temporal sparsity could be eventually revealed given a long observation period such as three weeks.

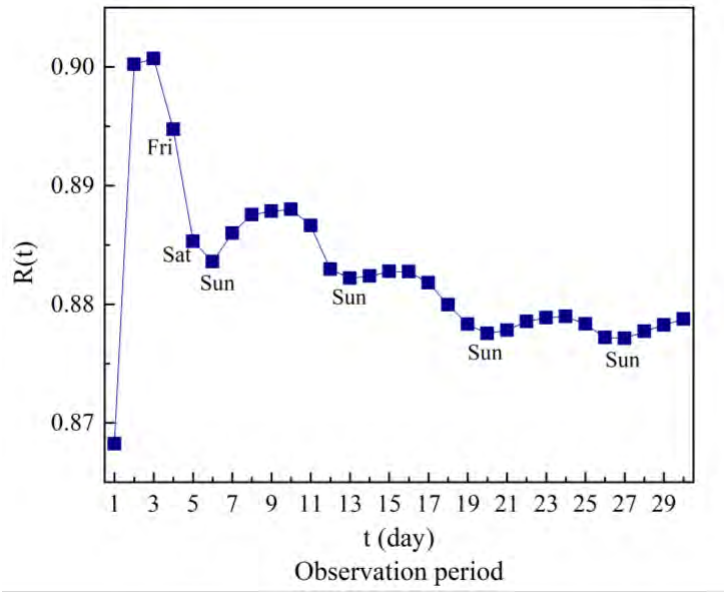


Figure 30. Graph. The observed regularity of visiting anchor locations R(t) when users are observed up to day t. (The first day April 1, 2014 is Tuesday).

Similarly, Figure 31 gives the evolution of mean radius of gyration r_g of activity locations that are observed up to time t (González et al. 2008). Here, radius of gyration of one user characterizes the traveled distance when the user is observed up to time *t*, and is calculated using the equation given in Figure 32, where L_i $(i = 1, \ldots, n(t))$ are activity locations observed and L_c gives their centers.

If r_g is stable, the inferred linear characteristic of users' trajectories is reliable only given a sufficient long observation period (e.g., longer than three weeks).

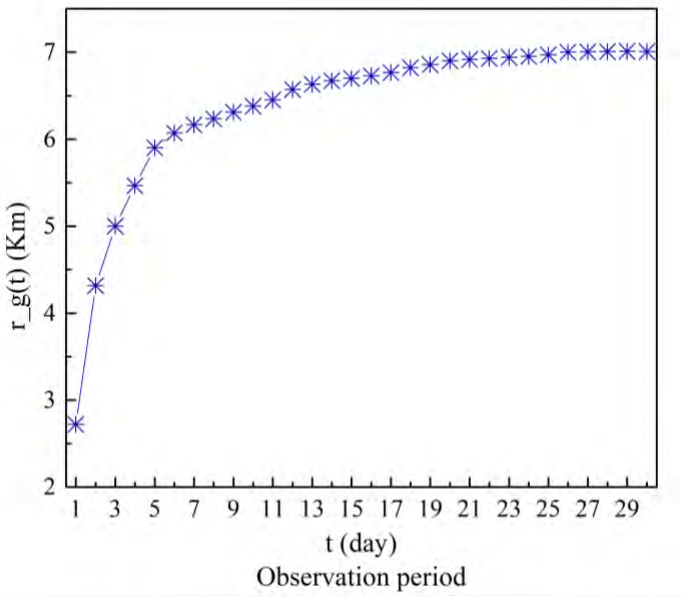


Figure 31. Graph. Evolution of radius of gyration r_g of activity locations with observation period.

$$r_g(t) = \sqrt{\frac{1}{n(t)}\sum_{i=1}^{n(t)}(L_{i-}L_c)^2}$$

Figure 32. Equation. Radius of gyration r_g of activity locations.

4.3 Second-order Properties

Once activity locations are identified, trips that connect two consecutive activity locations or trip ends are inferred. Then, individual derived trips, with origination and destination available for each trip, can be aggregated to estimate travel demand for a region.

4.3.1 Distribution of Trip Rates

Figure 33 gives the distribution of trip rates derived from the processed MPD. 91% users are observed with no more than four trips per weekday, leading to a mean of 1.8, which is significantly less than the observation from the 2002 travel survey—3.9 trips on average (Figure 34). This suggests that some of the trips are not captured by the MPD due to the temporal sparsity. The distribution of weekend trip rates is similar, with a mean of 1.6.

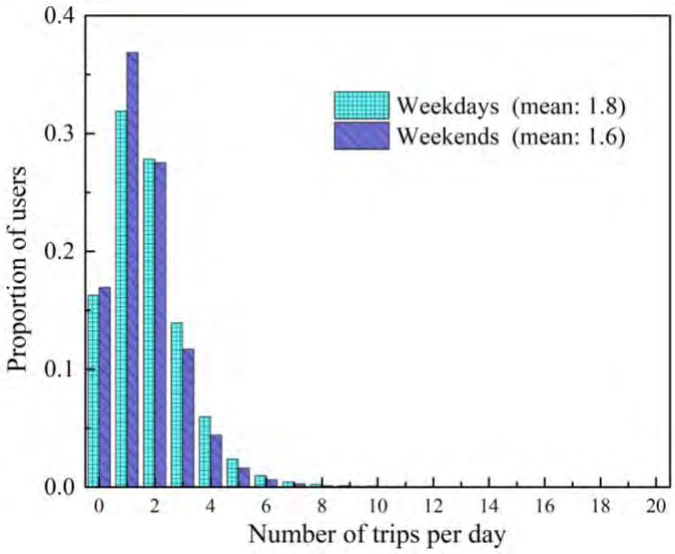


Figure 33. Graph. Distribution of trip rates on weekdays and on weekends.

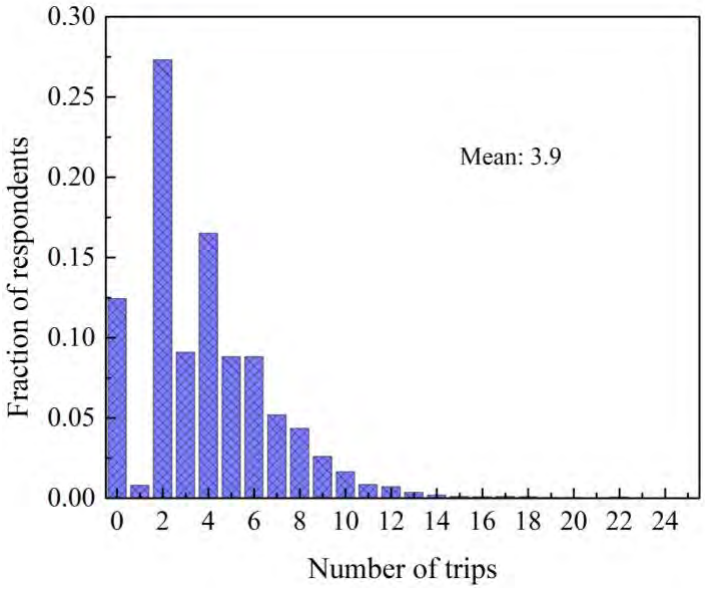


Figure 34. Graph. Distribution of trip rates from Buffalo travel survey.

4.3.2 Effect of Temporal Sparsity on Detecting Trip Rates

Figure 35 shows the correlation between the temporal sparsity of trajectories and the trip rates derived. Trajectories are divided into 24 groups according to their temporal resolution (the number of hourly slots with at least one sighting) and the average number of trips per trajectory is analyzed for each group. A clear correlation can be observed: the more severe the presence of temporal sparsity is, the fewer trips are detected. It suggests that trip rates could be underestimated due to temporal sparsity. This is not saying that the estimation of trip rate will be more biased if the temporal resolution is lower, as it is possible that users who make less calls conduct less travel (Iovan et al. 2013; Yuan, Raubal, & Liu, 2012). The number of trips observed becomes flat when the temporal resolution of trajectories φ reaches 16, suggesting that it is likely that no trip is conducted in the other remaining eight hours, which are possibly spent sleeping.

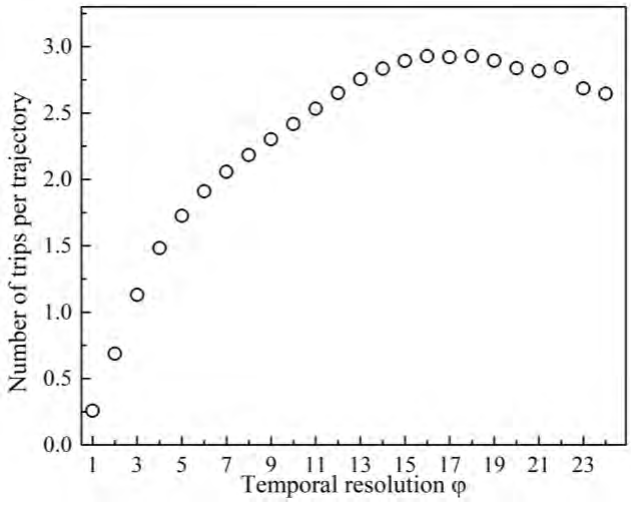


Figure 35. Graph. Correlation between temporal sparsity and derived trip rate.

In order to show more clearly how the temporal sparsity is related to the underestimation of the trip rate, the distribution of departure time of weekday trips derived from MPD is compared with what is observed from the travel survey data (Figure 36). Two peaks (morning and afternoon peaks) are clearly observed from the survey data. However, due to the low phone usage during the morning hours (see Figure 7), the MPD cannot adequately capture trips conducted during the morning rush hours. The absence of morning peak, which accounts for a large proportion of travel demand, consequently leads to the underestimation of the trip rate. The same pattern can also be observed in Figure 37, where the distribution of arrival time from MPD is compared against that from the Buffalo household travel survey data.

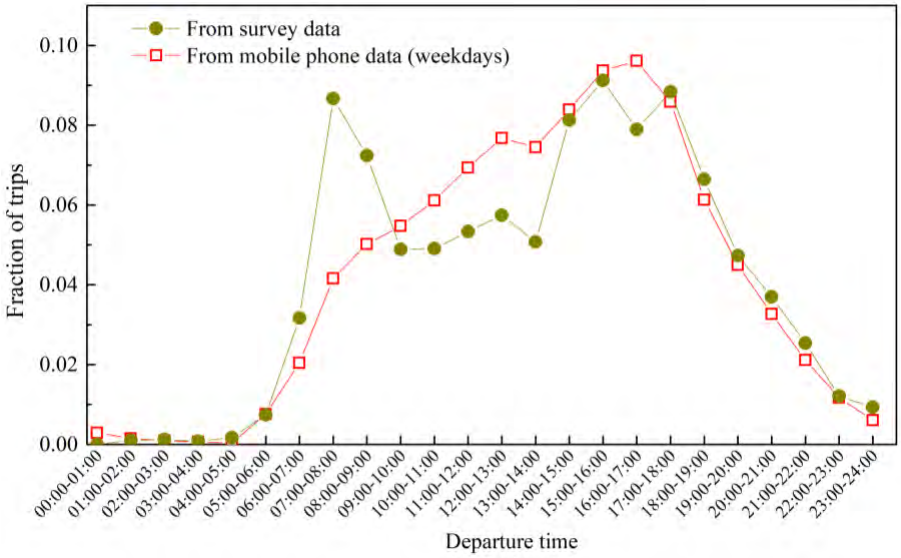


Figure 36. Graph. Comparison of distribution of departure time derived from MPD with that from the travel survey.

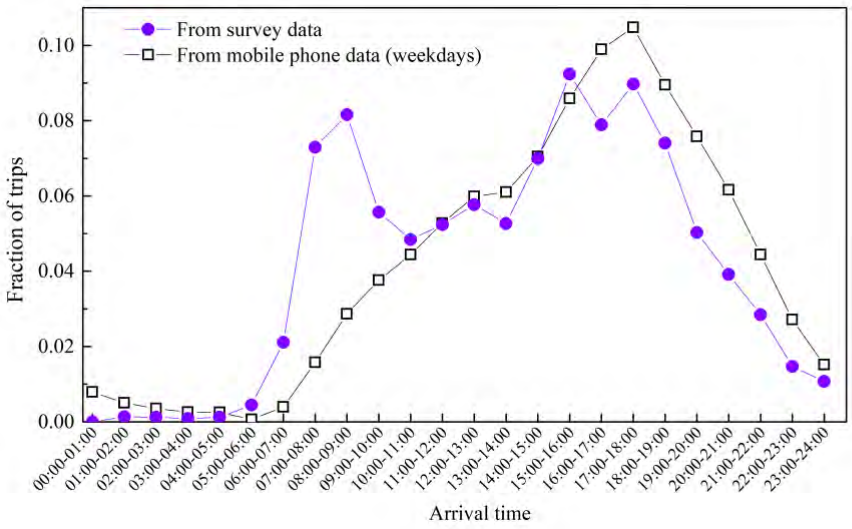


Figure 37. Graph. Comparison of distribution of arrival time derived from MPD with that from the travel survey.

4.3.3 Effect of temporal sparsity on capturing trip time

Besides the trip rate, the estimation of departure and arrival time of trips are also affected by the temporal sparsity. As Figure 38 demonstrates, the observed time departing *activity location 1* and arriving at *activity location 2* deviates from the actual times, leading to overestimation of trip time. This problem becomes more acute if certain trips are not captured at all.

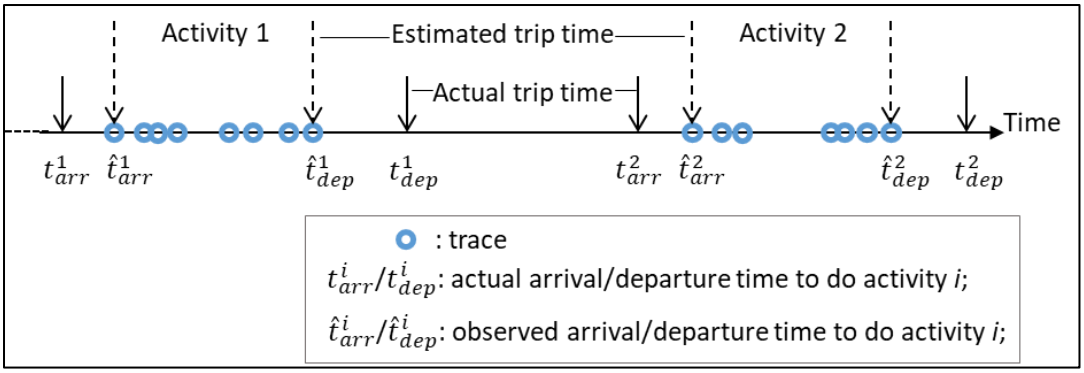
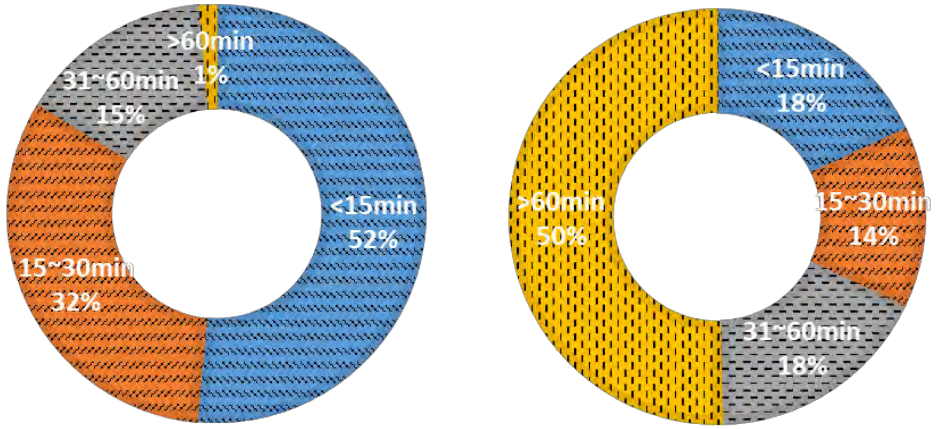


Figure 38. Graph. Demonstration on the difference between the estimated and actual trip time.

Figure 39 illustrates this issue by comparing the distributions of trip times derived from survey data (Figure 39-A) and from mobile phone data (Figure 39-B). Overall, trips recorded in the travel survey were of shorter duration. Half (50%) are less than 15 minutes in length, while another 32% are less than 30 minutes in length. Only 2% of trips are longer than 60 minutes. Observations from MPD, however, show a different picture: trips are inferred to have significantly longer durations with half of them exceeding one hour. Again, this is a phenomenon likely caused by temporal sparsity as noted earlier.



A. Trip time derived from survey data.

B. Trip time derived from mobile phone data.

Figure 39. Graph. Comparison of trip time derived from mobile phone data and from survey data.

4.3.4 Effect of Oscillation on Detecting Trips

Besides temporal sparsity, the issue of oscillation, if not properly addressed, will also alter the estimation of trip rate. Figure 40 gives the distribution of trip rates derived from the data without

removing oscillation traces, showing a higher mean value (2.7) than the one observed from the processed data (1.8) (see Appendix A. Processing Mobile Phone Data on how to detect and remove them). The distribution also indicates there is a non-negligible proportion (5%) of users conducting more than 10 trips per day, which is much higher than the 3% observed from travel surveys. These observations suggest that the trip rates tend to be overestimated without removing oscillation traces.

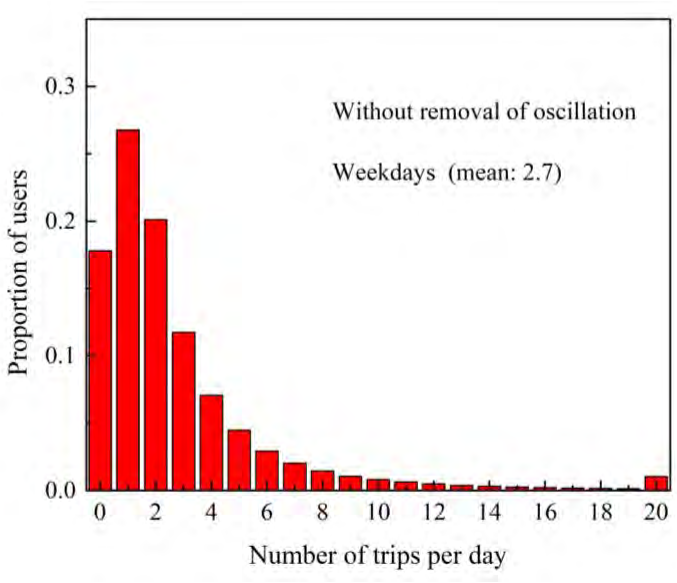


Figure 40. Graph. Distribution of trip rate without removing oscillation traces.

4.3.5 Weekly Pattern of Trip Rates

The travel survey data is a snapshot of travel demand at a fixed time. However, the longitudinal nature of MPD enables one to capture variations in travel demand over time. Figure 41 gives the daily trip rate per user during the study period. It shows a clear weekly pattern, with the highest demand on Fridays and lowest demand on weekends (especially on Sundays). Though it is intuitive that people travel less on weekends, the relatively sparser data on weekends (see Figure 10) also contributes to the decline of observed trips. Note that unexpected drops on the first and the last day may be due to the first and last days of the study period.

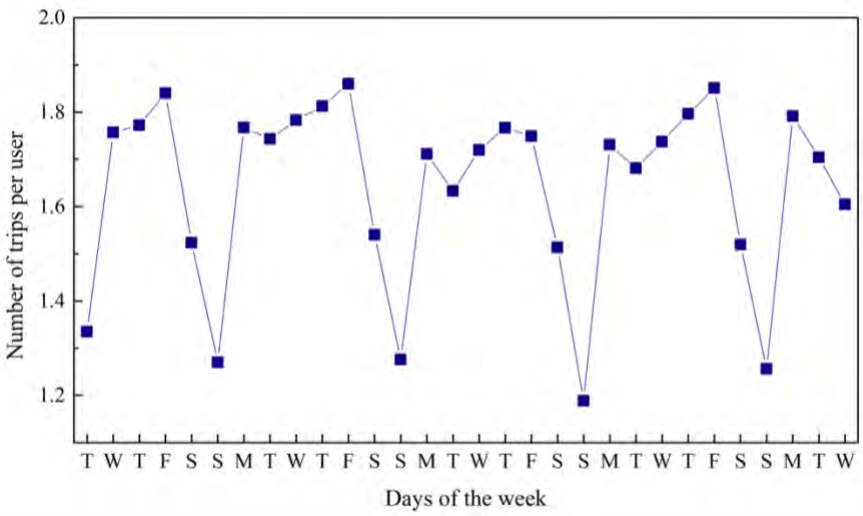


Figure 41. Graph. Weekly pattern of trip rate detected.

The distribution of departure times of weekday trips derived from MPD is compared with that of weekend trips (Figure 42). It shows clear differences in travel demand between weekdays and weekends. Though the absence of weekday morning peak is due to low phone usage in the morning, the weekday late-afternoon peak is observed. However, the peak of travel demand on weekends appears at midday period. This suggests there is no late-afternoon peak on weekends. Given the low phone usage during the morning hours of weekends, no conclusion could be drawn on the existence of the morning peak on weekends. The distribution of arrival time shows similar patterns but with peaks later than those of departure time (Figure 43).

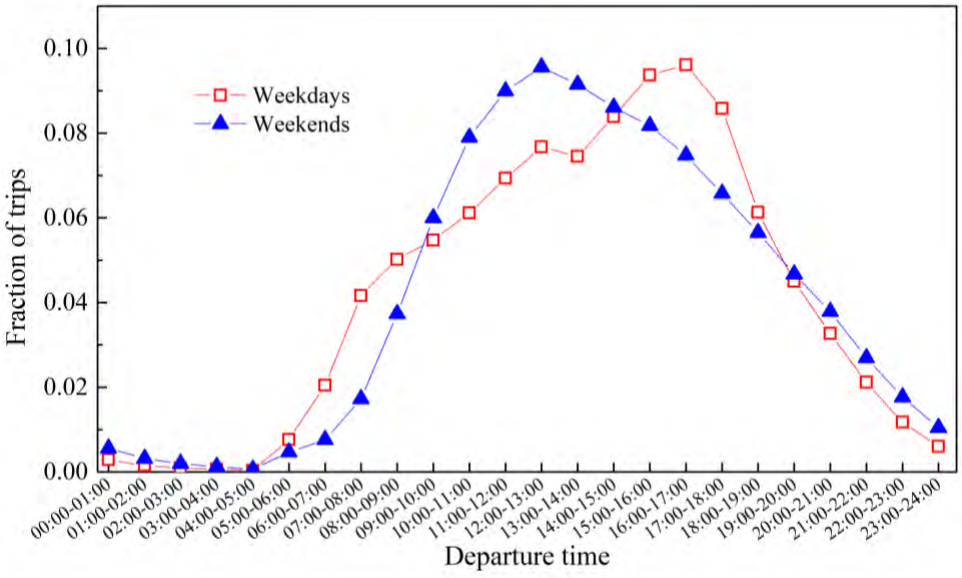


Figure 42. Graph. Comparison of distribution of departure time of weekday and weekend trips derived from MPD.

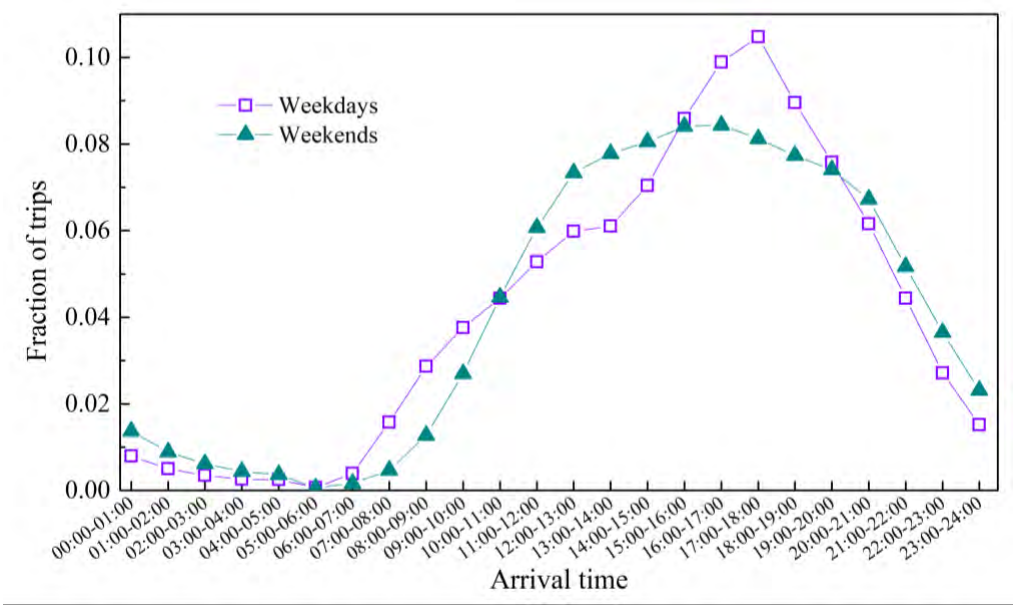
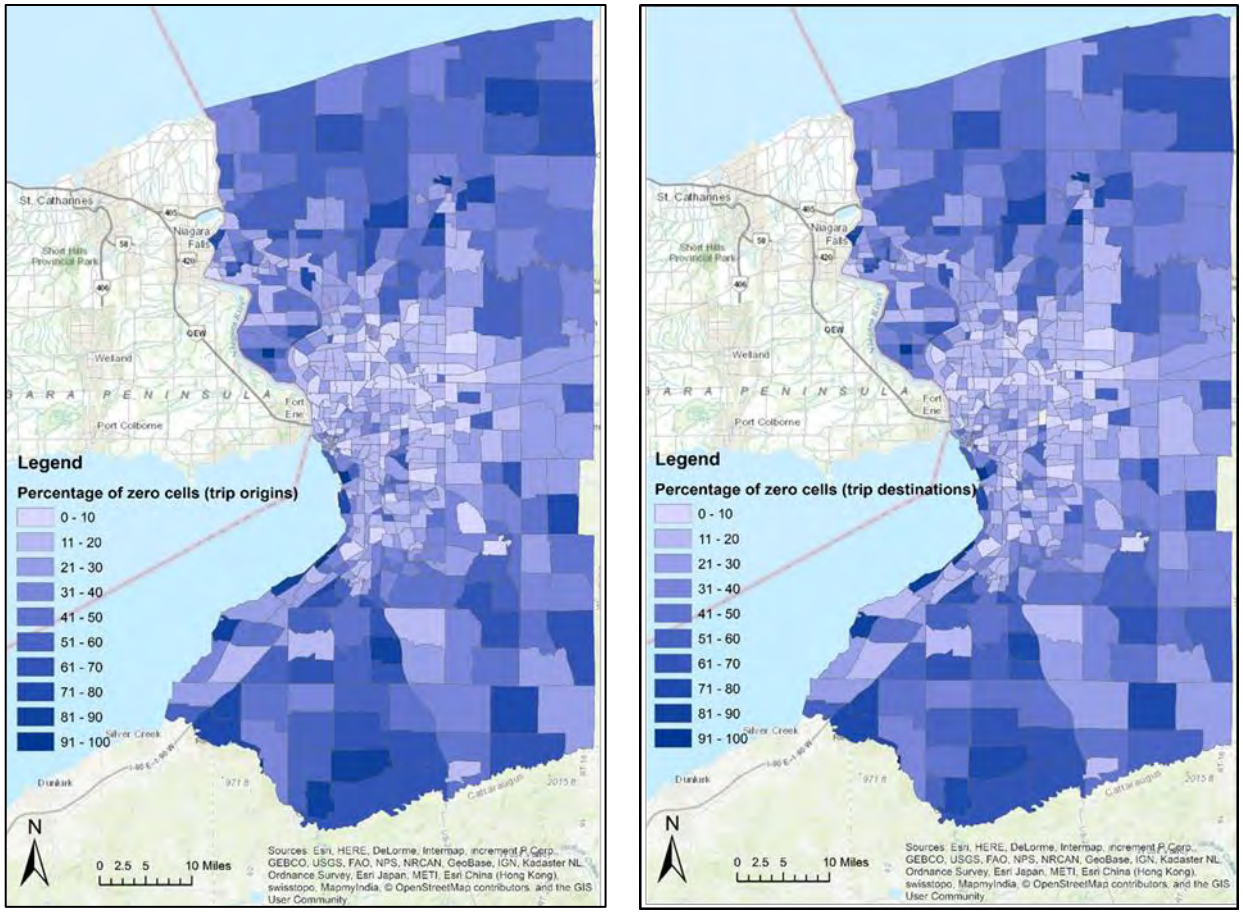


Figure 43. Graph. Comparison of distribution of arrival time of weekday and weekend trips derived from MPD.

4.3.6 Proportion of Zero Cells in Origin-Destination Matrix

In an origin-destination (OD) matrix, each cell represents the number of trips from one traffic analysis zone (TAZ) zone to another TAZ zone, and intrazonal trips are represented in the diagonal cells in the matrix. For each TAZ, the percentage of cells (TAZs) with no trip originating from or destining to that particular TAZ is calculated among all TAZs. They are shown as percentage of zero cells of trip origins in Figure 44-A and of trip destinations in Figure 44-B. These percentages, either in terms of trip origins or destinations, are low in the downtown area (mostly lower than 20%), but are relatively high in the rest of the study area (higher than 50% mostly), especially further south and north zones from downtown (range between 70-100%). This indicates that it is possible to observe at least one trip from/to most zones in the downtown Buffalo area. Overall, the two distributions are similar.

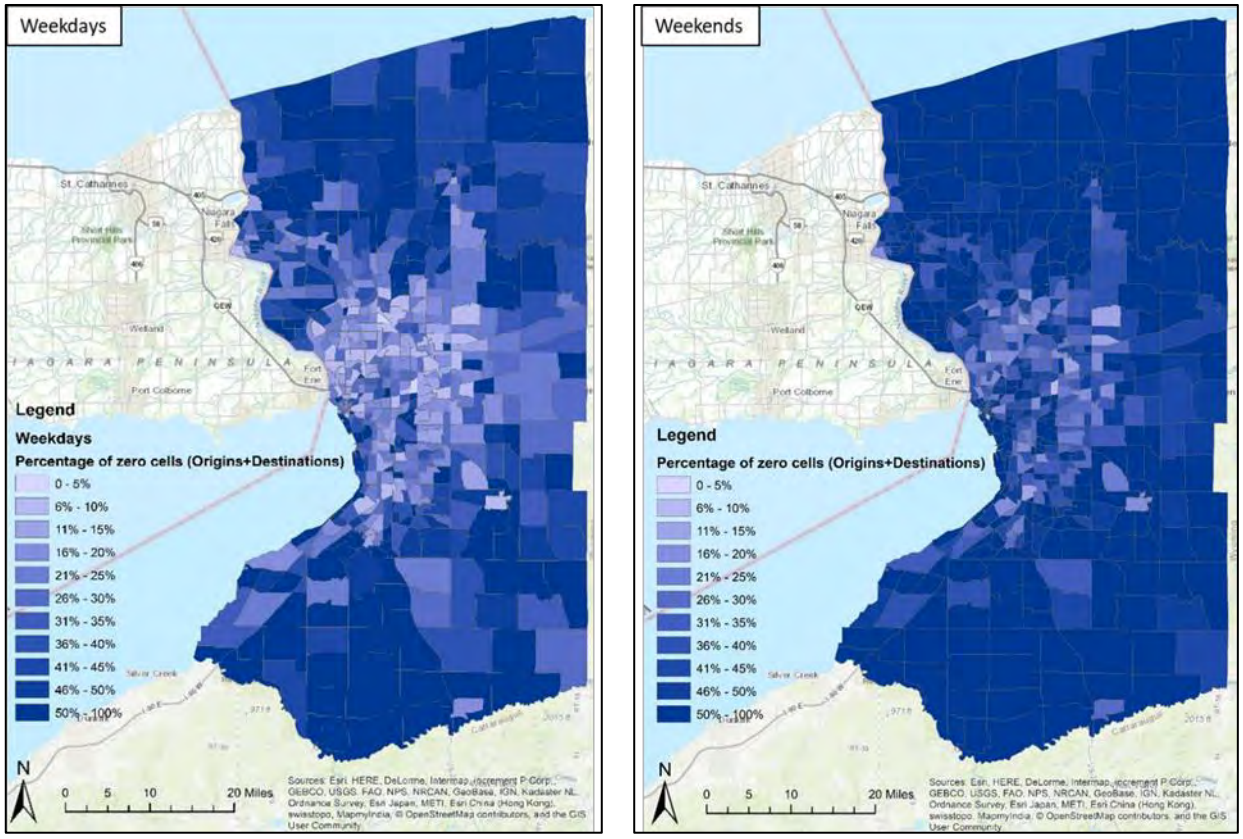


A. Zero cells of trip origins. B. Zero cells of trip destinations.

Figure 44. Graph. Comparison of spatial distributions of zero cells of trip origins (A) and trip destinations (B).

Source: © ESRI, World Topographic Map

Additionally, for each TAZ, the percentage of cells (TAZs) with no trip originating from and destining to that particular TAZ is also calculated among all TAZs. They are shown as percentage of zero cells (trip origins and destinations) in Figure 45 (weekdays on the left and weekends on the right). During weekdays, most zones in the downtown area have less than 20% of zero cells whereas the zones in the rest of the study area have a much larger number of zero cells, mostly more than 50%. However, during weekends, most of the zones in the downtown area appear to contain more than 50% of zero cells.



A. Zero cells in weekday OD matrix.

B. Zero cells in weekend OD matrix.

Figure 45. Graph. Comparison of spatial distributions of zero cells (both trip origins and destinations) in weekday and weekend OD matrix.

Source: © ESRI, World Topographic Map

The daily variation of the percentage of zero cells in OD matrices is given in Figure 46. A weekly pattern is found: weekdays have lower percentage of zero cells than weekends, and the lowest and the highest percentages of zero cells are found on Friday and Sundays, respectively. This observation is consistent with previous findings that Fridays have the most number of sightings and Sundays have the least.

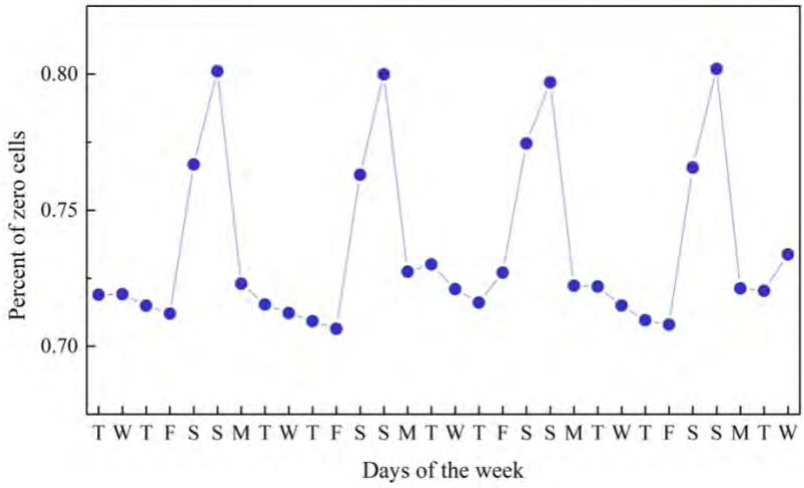


Figure 46. Graph. Daily variation of proportion of zero cells.

Figure 47 shows how the percent of zero cells changes with daily accumulation of data. The percent of zero cells is quickly reduced in the first week of the study period (the accumulation of 1-7 days of MPD quickly reduce zero cells in OD matrix), but the speed of the reduction is substantially decreased in the second week and then stabilizes. On the first day of observations, about 70% of cells are zero, but on the third and sixth days, about 60% and 50% of zero cells remain. In the third and fourth weeks, the proportions of zero cells are between 30-40%. This finding is also consistent with the earlier finding that 3 weeks appear to be a reasonable time frame to capture most of the activity locations visited (Figure 30).

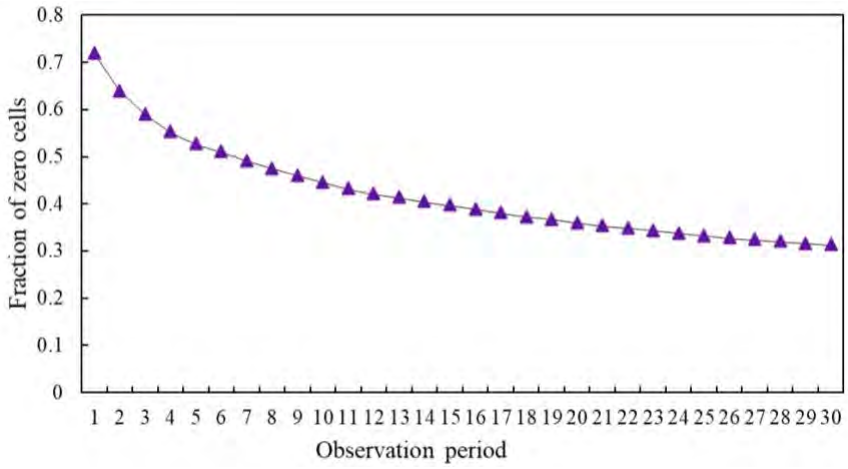


Figure 47. Graph. The decrease of the percent of zero cells with accumulation of data.

4.3.7 OD Collinearity and Compared with MPO OD

Because the MPD does not represent the entire population of the region, and because the total number of trips extracted from this data only account for about 18–20% of the total trips found for the region based on MPO model results, it is thus necessary to scale up the extracted OD trips. The scaling factors are calculated by dividing the population by the number of inferred homes in each TAZ zone. In this way, the OD trip can better represent population and adjust the difference between population distribution and inferred homes from the MPD (see Section 3.1). However, if

the number of inferred homes is zero, the value after scaling is also zero. The mean of scaling factors is 3.5 and the first, second, and third quartile values are 0.63, 1.28, 2.34, respectively. Figure 48 shows the collinearity between the OD matrices extracted from MPD and the OD matrix before and after the scaling-up with the scaling factors. Although the correlation coefficient is 0.864, the number of trips expanded from the upscaling process only increases about 8.3% from 18% to 26.3% of the total number of trips from the MPO model results. In total, as a result, about 16 million OD trips are obtained from MPD and their similarities (TAZ level) to the MPO model results are about 0.66 (Figure 49).

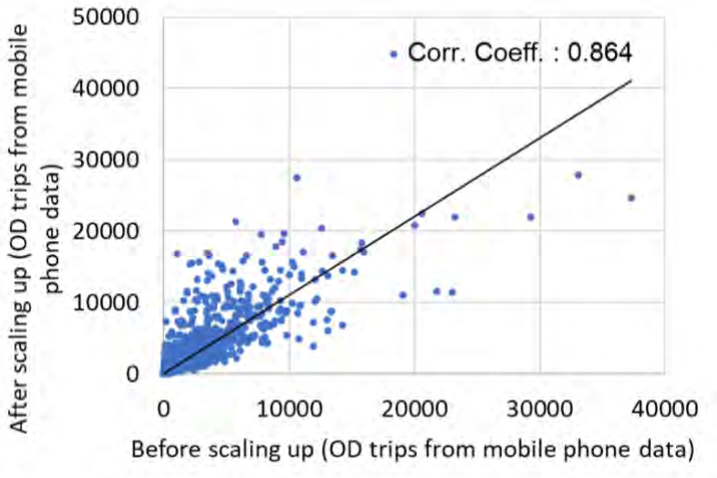


Figure 48. Graph. OD Collinearity.

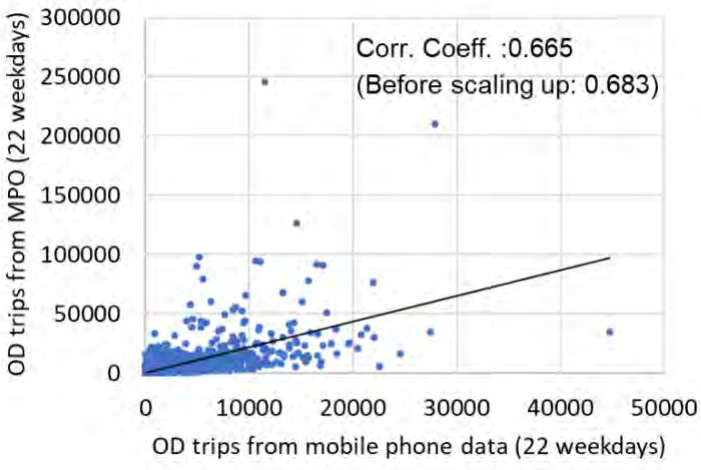


Figure 49. Graph. Comparison between up-scaled OD matrix from mobile phone data and MPO OD matrix.

4.3.8 OD Sensitivity Analysis

One of the main issues of using all users in this analysis is the uncertainty of home location inference for users without enough records. These users provide insufficient records to extract/estimate their daily trips, and as a result, there may be underestimation of OD trips in the scaling-up process described in Section 4.3.7. For example, while the number of users in the

MPD is about 80% of the population in this area (933,508 users / 1,170,111 people), the expanded number of OD trips only account for about 26% of OD trips from MPO results. Because of this issue, Alexander et al. (2015) also used a filter to select users who provide enough records to infer their home locations. Their filter is at least one visit at each user's inferred home in a week (4 visits in a month), resulting in about 335,795 users used in their analysis, which accounts for about 10% of population in Boston,[6] where the study was conducted. This filter was applied in this report's OD sensitivity analysis and was termed Alexander's rule. Additionally, various combinations of different filters were tested. In general, two sets of filters are found: one is a lenient rule (at least 3 nights and 3 days in a month, and average three hourly time-slots and three sightings in a day), and another is a strict rule (at least 6 nights and 9 days in a month, and average five hourly time-slots and four sightings in a day).

Table 9 shows the statistics of scaling factors using different rules. The scaling factors are sensitive to the different rules applied in the analysis. For example, the medians of the scaling factors range between 1.28 and 7.94, and the third quartiles of scaling factors are between 2.34 and 15.37.

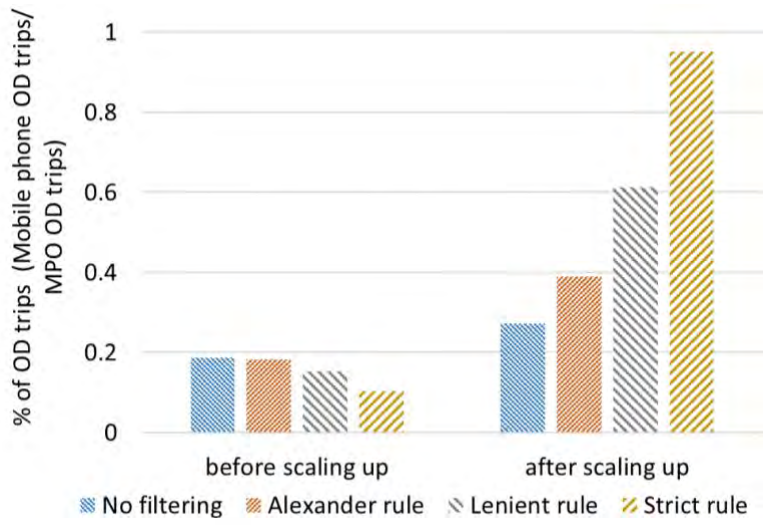
Table 9. Scaling factors using different filters.

	No filtering	**Alexander rule**	**lenient rule (3,3,3,3)**	**Strict rule* (6,9,5,4)**
1st quartile	0.6341	1.034	1.823	4.231
median	1.2822	1.861	3.666	7.936
3rd quartile	2.3387	3.530	6.798	15.368

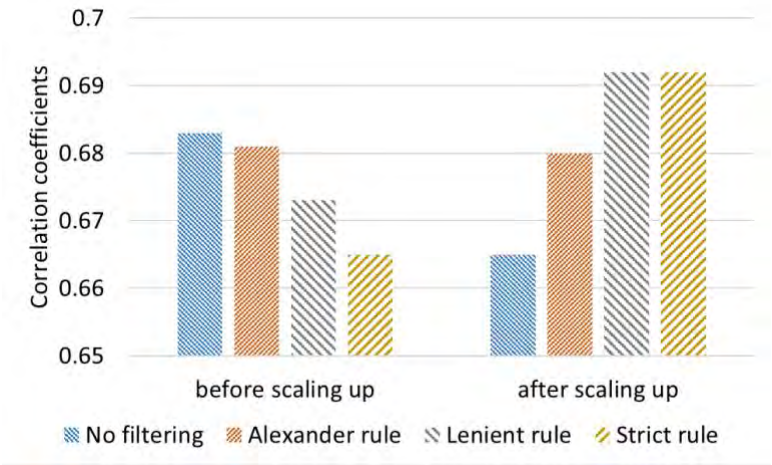
As a result, the total number of OD trips expanded with these scaling factors are also different with each other (Figure 50-A). Before upscaling, the total number of OD trips only accounts for less than 20% of OD trips in MPO model results. However, it is possible to estimate about 26–95% of OD trips in MPO model results after the upscaling process.

The correlation coefficients between the OD trips from MPD and MPO results before-and-after upscaling process are provided in Figure 50-B. The coefficient after the upscaling process is decreased by 0.02 when no-filter is applied. Meanwhile, the coefficients are almost the same before-and-after upscaling when Alexander's rule is applied. However, when the lenient and strict rules are applied, the similarity between inferred trips from MPD and MPO results increases by 0.03. Overall, applying the strict rule seems to produce the best results in terms of the similarity between the OD matrices derived from the MPD and the MPO results.

[6] This is the most recent study and the only study area in the US that OD matrix were scaled up based on the comparison between population and inferred home location.



A. The ratio of the number of trips derived from mobile phone data and that from MPO data.



B. Correlation coefficients between OD matrixes derived from mobile phone data and from MPO data.

Figure 50. Graph. OD sensitivity analysis of filters.

5.0 GPS Data

The GPS dataset was provided by a data vendor who integrated GPS data from about 15 raw data sources (e.g., fleet companies and mobile apps), which encompasses 90 days for part of the City of Seattle. For each vehicle, its vehicle ID (VID), coordinates of location, speed, and time were provided.

Figure 51 shows a map of the study area and the GPS data points from one day.

The study area of GPS data is relatively small, and the trips and their characteristics focus more on short trips (less than 30 minutes in most cases). They present different patterns compared with the mobile phone data that cover an entire metropolitan area. Throughout this chapter, comparisons are made for such differences when appropriate. The patterns and findings here from the GPS data may not be applied directly to GPS data for a much larger area such as a region. However, it is understood that the analysis framework and some of the methods (such as trip end identification) may be applied to GPS data on different sizes of the study areas.

Similar to the mobile phone data, an important temporal feature of the GPS data is the different pattern for weekdays and weekends, as most people need to adjust their travel schedules due to working hours on weekdays but do not need to do so on weekends in most cases. Therefore, temporal characteristics are analyzed separately for weekdays and weekends and for the overall dataset. Daily, weekly, and monthly analyses are also conducted to see how the results change for different lengths of the analysis periods.

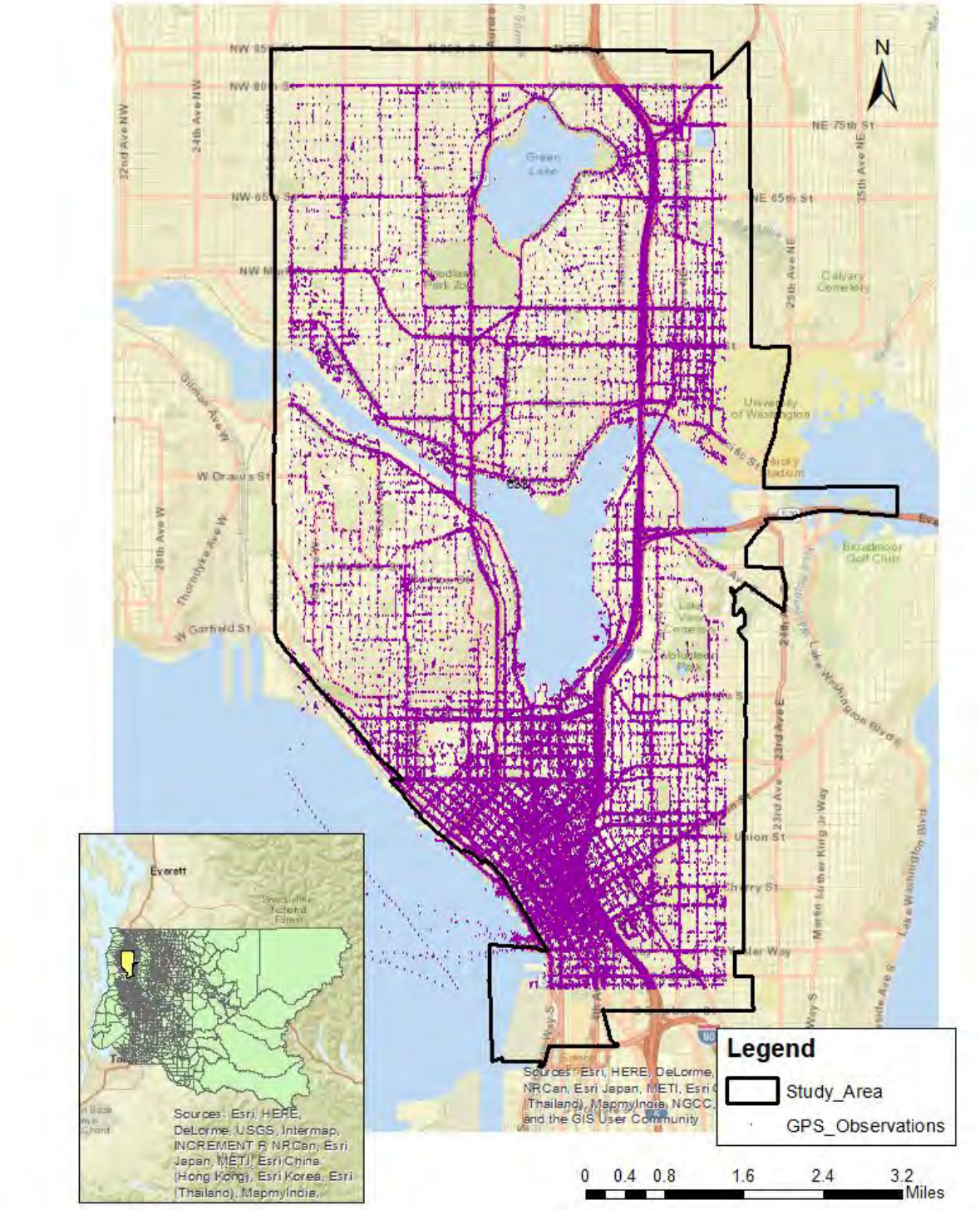


Figure 51. Map. The study area and raw GPS observations from one day.

Source: © ESRI, World Street Map

5.1 Zeroth-order Properties

The zeroth-order properties aim to capture the overall characteristics of the GPS data by analyzing characteristics such as the lifespan of VIDs, sampling interval, temporal pattern of the observations, as summarized in the following subsections.

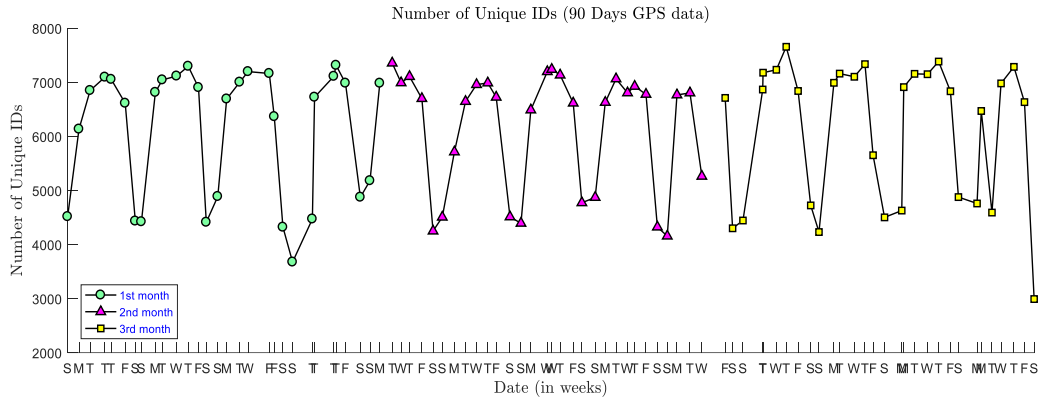
5.1.1 Vehicle ID Analysis

VIDs were randomly generated to protect the vehicle privacy. The GPS data used in this project were collected from multiple data sources (about 15 or so), each with different policies and algorithms to generate VIDs. Some update VIDs for the same vehicle every a few hours (often two hours), some update them every day, and some never update the VIDs. Therefore, while the same VID represents the same vehicle in the dataset, a single vehicle may be represented by multiple VIDs, depending on the specific VID generation algorithm. Unfortunately, there is no way (even for the data vendor) to know exactly how the VIDs were generated and how frequent they were updated in the GPS dataset. This is different from the mobile phone data for which the device IDs did not change over the study period.

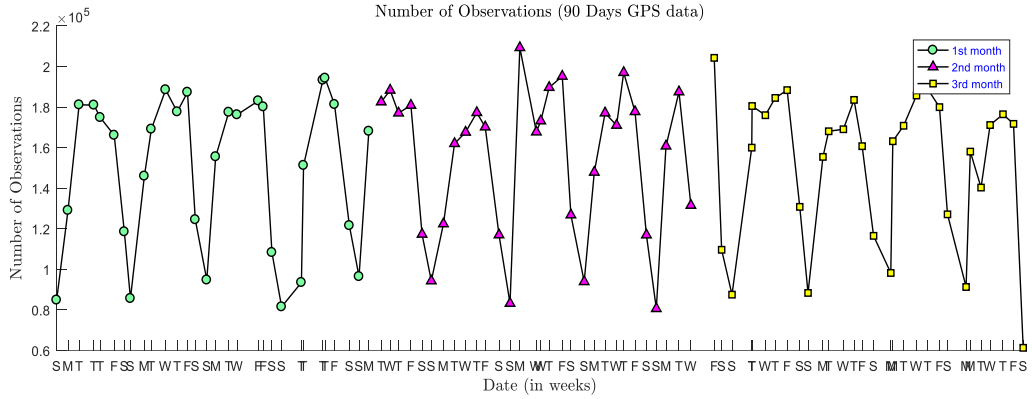
In this subsection, the number of observations and unique VIDs are summarized to obtain a basic understanding of the GPS data. From Figure 52-A and B, the number of observations and the unique VIDs on weekdays are larger than those on weekends. The numbers of unique VIDs on Sundays and Saturdays are relatively close, while the number of observations on Sundays are usually much smaller than that on Saturdays. From Figure 52-C, it is found that in general, the number of unique VIDs takes up 3.5%~5.5% of the number observations. The ratio of the number of unique VIDs over the number of observations shows a daily fluctuation pattern and always peaks on Sunday, which may be caused by the lowest number of observations on Sundays. Table 10 summarizes the number of observations and unique VIDs for each month. The ratio of the number of unique VIDs over the number of observations is around 4% in every month.

Table 10. The number of observations and the number of unique VIDs (monthly).

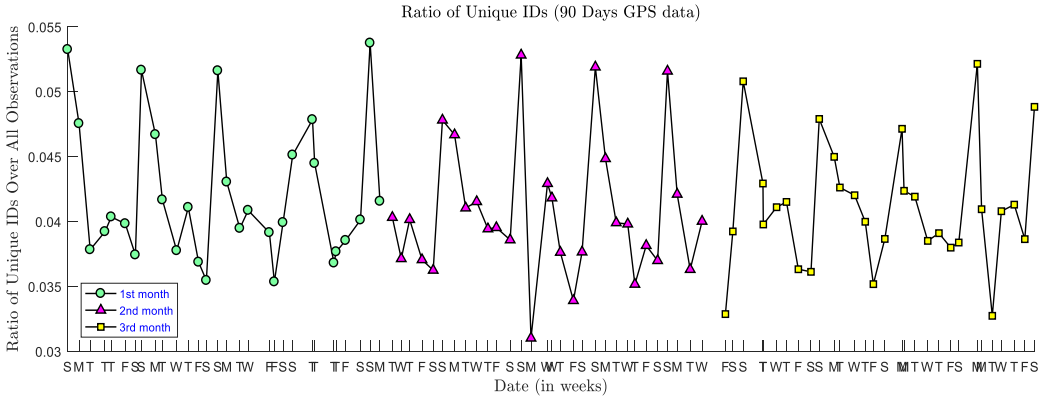
	Number of Unique VIDs	Number of Observations	Unique VIDs Ratio
The first month	183,643	4,470,254	4.11%
The second month	184,654	4,643,150	3.98%
The third month	183,563	4,547,035	4.04%



A. The number of unique VIDs (daily).



B. The number of observations (daily).



C. Unique VIDs ratio (daily).

Figure 52. Graph. The number of observations and the number of unique VIDs (daily).

To further investigate the pattern of the VIDs, the term “lifespan” of a unique VID is defined as the time difference between the first time and the last time when the VID was observed. The term indicates for how long a particular VID may have existed in the study area as reflected in the GPS data. Notice that, if VIDs are updated frequently, the lifespan of a VID will not necessarily correspond well to the actual duration that the vehicle has existed in the area. Figure 53 shows the histogram of VID lifespans (measured in number of days) for each month separately. It is obvious that most VIDs (nearly 90%) only existed for one day in the area. However, there are a small portion of VIDs that existed for multiple days and even for the entire month. The reasons

for this VID lifespan pattern could be: (i) some vehicles, in particular trucks, only passed through the study area once and the VID observations reflect the true pattern of those trips; (ii) the VIDs may have changed every day. Both cases may exist in the dataset, but there is no way to figure out which of the two cases is the most dominant.

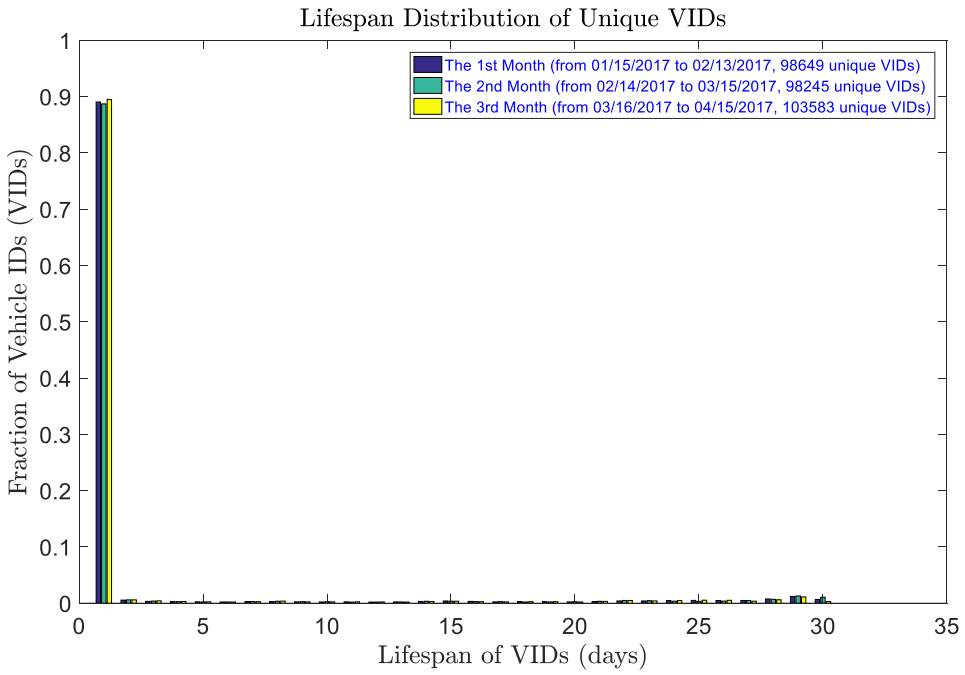


Figure 53. Graph. Lifespan distribution of all unique VIDs (monthly data).

Since the lifespans of most VIDs are less than one day, the lifespan distribution of VIDs within a day is shown in Figure 54 (in log-scale). Out of the VIDs whose lifespans are within one day, most VIDs (nearly 90%) only lasted for less than 1 hour. This is intuitively understandable because the study area is relatively small. It takes less than an hour for a vehicle to complete its activities through the study area, unless the vehicle stays in the area for multiple trips (activities).

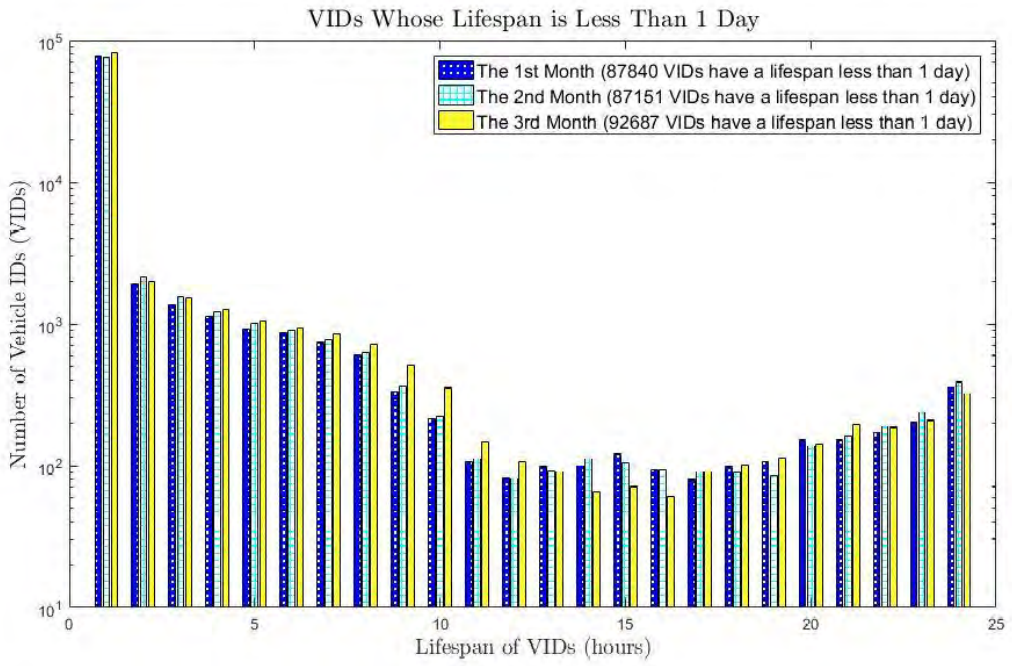


Figure 54. Graph. Lifespan distribution of VIDs within a day (monthly data).

As around 80% of VIDs have a lifespan less than 1 hour. Figure 55 shows the VIDs whose lifespans are less than 1 hour. Most VIDs appeared for 30 minutes or less.

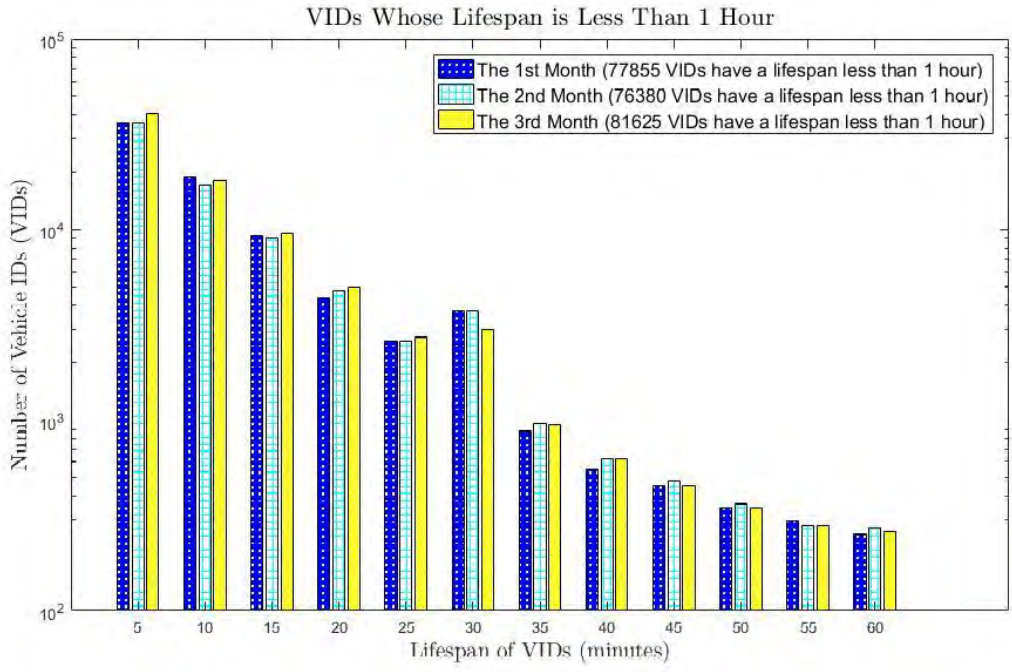


Figure 55. Graph. Lifespan distribution of VIDs within 1 hour (monthly data).

5.1.2 Sampling Interval

The sampling interval indicates the time difference between two consecutive data records of the same vehicle (more accurately, of the same VID in the GPS dataset). It is an important parameter that determines the level of detail the collected data may have captured temporally. For each VID and within a trip (the trip identification method is discussed later), the time difference between two consecutive observations is computed and the distribution of the observation intervals can be obtained. Figure 56 and Figure 57 show the distributions of the sampling intervals for weekdays and weekends, respectively. Most GPS data sampling intervals are less than 300 seconds, and the most frequently utilized sampling intervals are 15, 60, 90, 18, 20, and 120 seconds. Comparing the histograms of weekdays and weekends, the sampling interval distributions are similar between weekdays and weekends. The distribution of GPS data sampling intervals (in log scales) are shown in Figure 58, similar to Figure 1 for the mobile phone data. The distribution of GPS data shows a nearly linear pattern, which is very different from that of the mobile phone data.

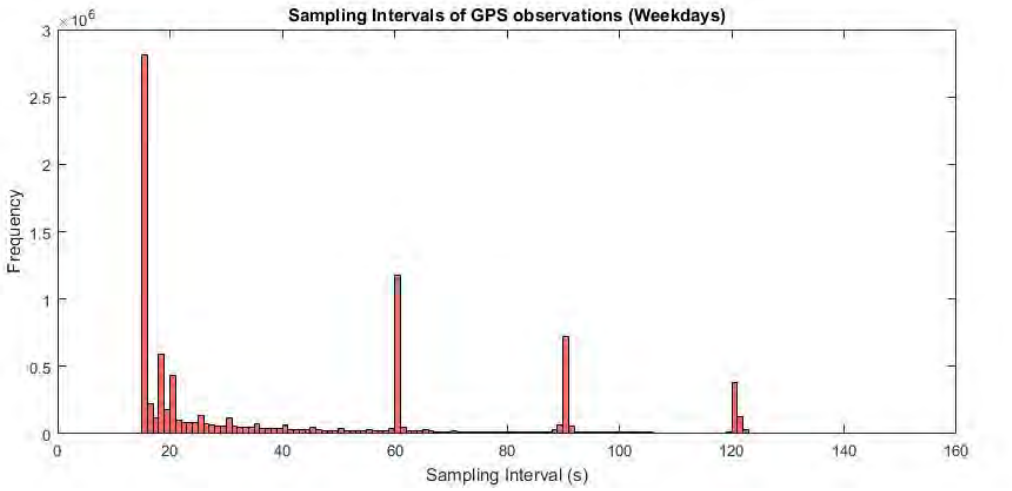


Figure 56. Graph. Distribution of the sampling intervals (all weekdays).

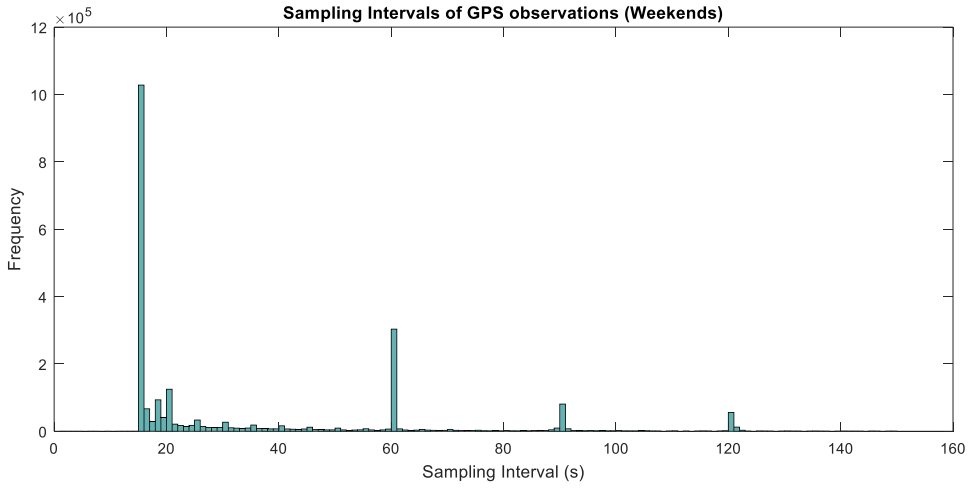


Figure 57. Graph. Distribution of the sampling intervals (all weekends).

In addition to the differences in the sampling intervals on weekdays and weekends, the distributions of sampling intervals for different lengths of the observation period (e.g., a week, a month, and three months) were also investigated. For various lengths of the period, it was found that the sampling interval distributions remain almost the same. This indicates that the GPS data sampling intervals are fairly stable and do not show obvious weekly or monthly variations.

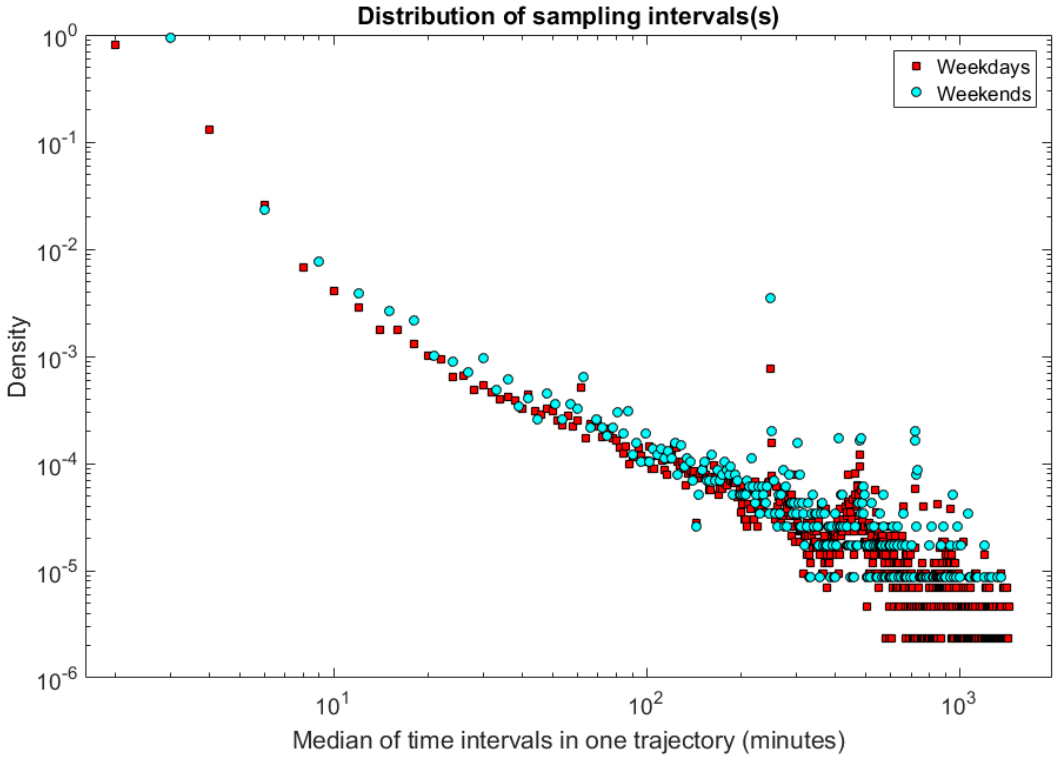


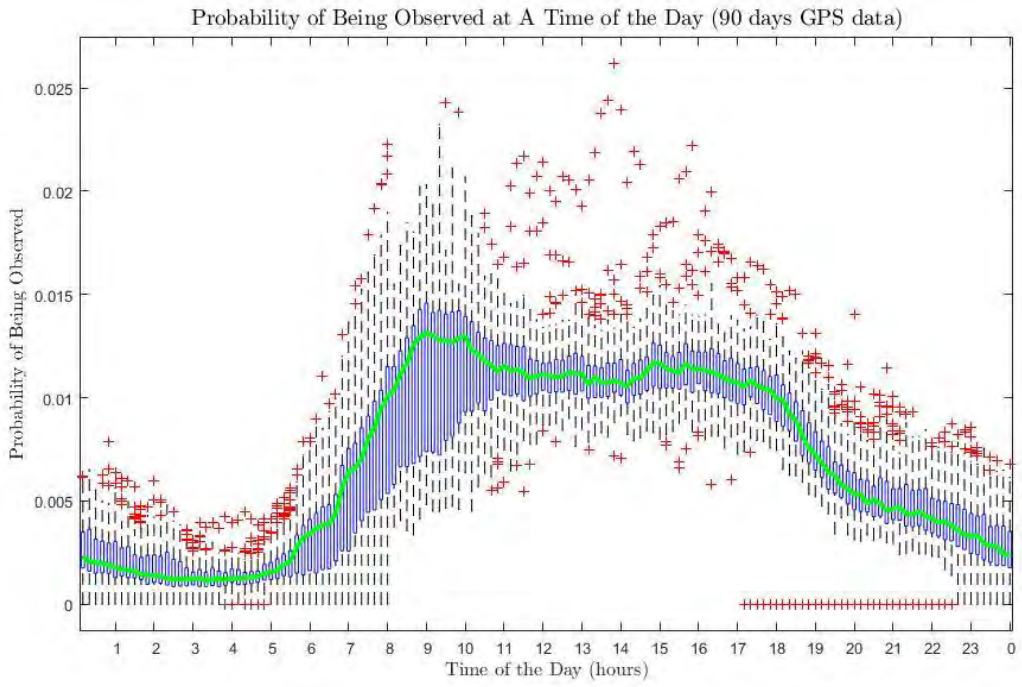
Figure 58. Graph. Distribution of the sampling intervals in log-scale.

5.1.3 Temporal Pattern of the Observations

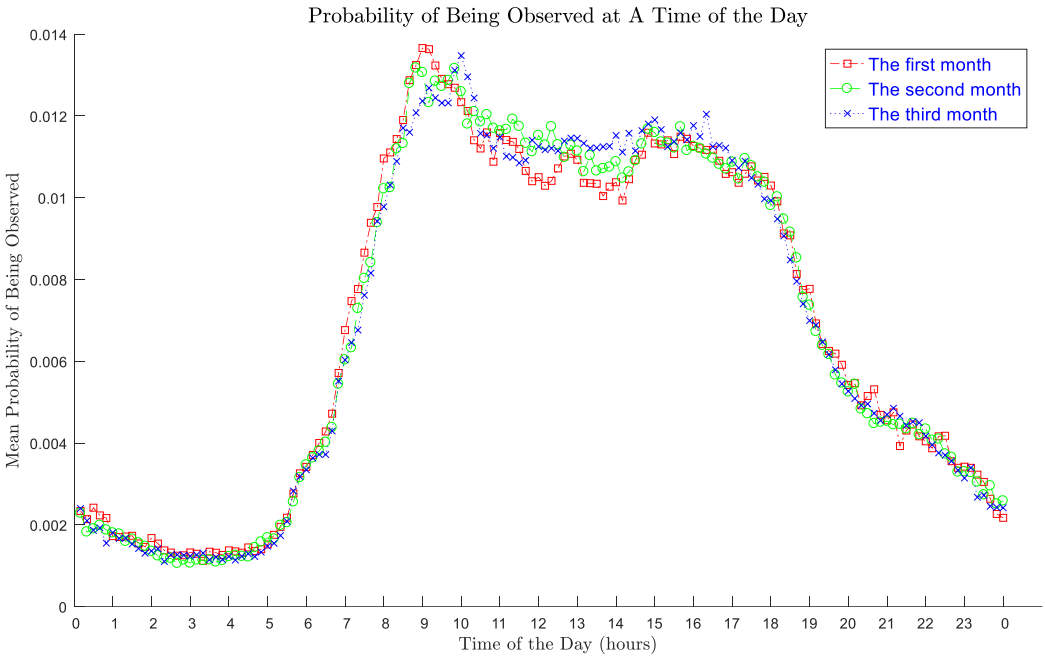
To study how the GPS data observations vary by time of a day, the time is first divided into 10-minute intervals. For a given day, the fraction of the number of observations of each 10-minute interval is calculated. The boxplot and the median (green curve) of each interval for the 90 days are presented in Figure 59-A. The monthly analysis results from Figure 59-B. show similar trends. The data show that the peak of the observations starts at around 7:00 and reaches the highest value at 9:00 which lasts for about an hour. Then it keeps at a relatively high and constant level between 10:00 and 17:00. Around 15:00, there is an afternoon peak and the number of observations begins to decrease starting at 17:00. It is intuitively understandable that a larger portion of the observations are for daytime, since most people are traveling or doing other travel related activities during the day. In addition, Figure 59-C and D, are patterns of the observations at the time of the day for weekdays and weekends, respectively. On weekdays, the frequency of observations starts to increase at around 5:00 and peaks at around 9:00, then the observations keep at a relatively high level through 17:00. For weekends, the fractions of observations in the morning have high variations, which implies that on some weekends there are large number of observations in the morning while this is not the case for other weekends. The observations on

weekends have a general trend of increasing from 6:00, and peak between 13:00 and 14:00. Besides, the pattern of observation of weekdays shows a relatively high value in the daytime, which indicates a high frequency of work-related activities, while on weekends it has a unimodal pattern that reflects more leisure activities and non-work-based trips.

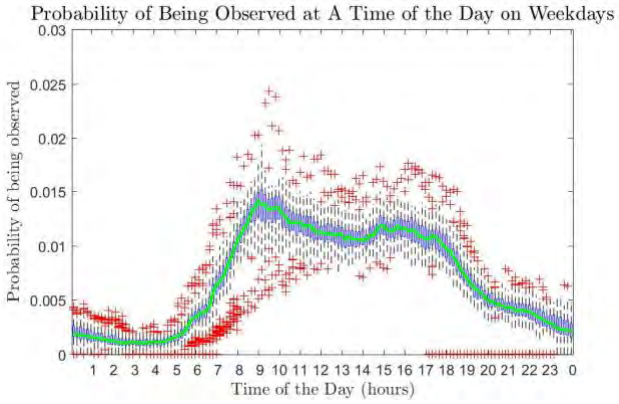
The findings from the GPS data indicate that the collected data have fairly strong correlation with vehicular travel patterns at the aggregated level. They may represent the morning/afternoon and weekend/weekday travel patterns reasonably well. This is quite different from what has been observed from the mobile phone data, as discussed in Section 4.1. The main reason for this is probably the fact that the GPS data were collected primarily from vehicles and thus are related more closely to vehicular travels, while this may not be the case for the mobile phone data.

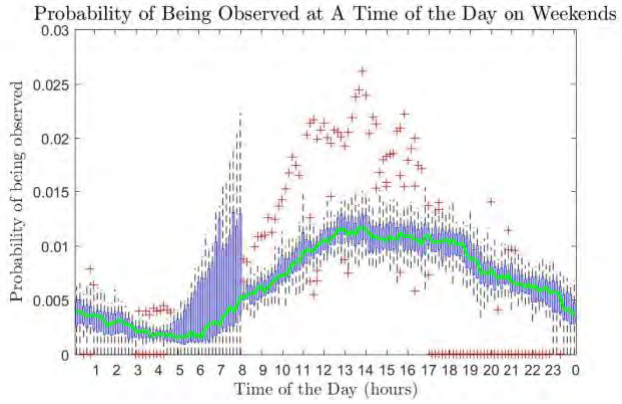


A. Probability of being observed.



B. Mean probability of being observed.





C. Probability of being observed on weekdays. D. Probability of being observed on weekends.

Figure 59. Graph. Pattern of observations.

5.1.4 Number of Days Observed

The number of days observed for a unique VID can provide information about whether a VID is recorded in multiple days, indicating how “active” the particular VID is during the study period and within the study area. Notice that this measure is different from the VID lifespan since it is possible that a VID lifespan is large (e.g., spans across multiple days), but the number of days observed is small (e.g., only appeared at the beginning and the end). In Figure 60, out of all the unique IDs of the 90-day data, 89.7% have been observed within only one day, 4.7% have been observed

for two days. This pattern indicates that the VIDs of GPS data have relatively low consistency from day to day.

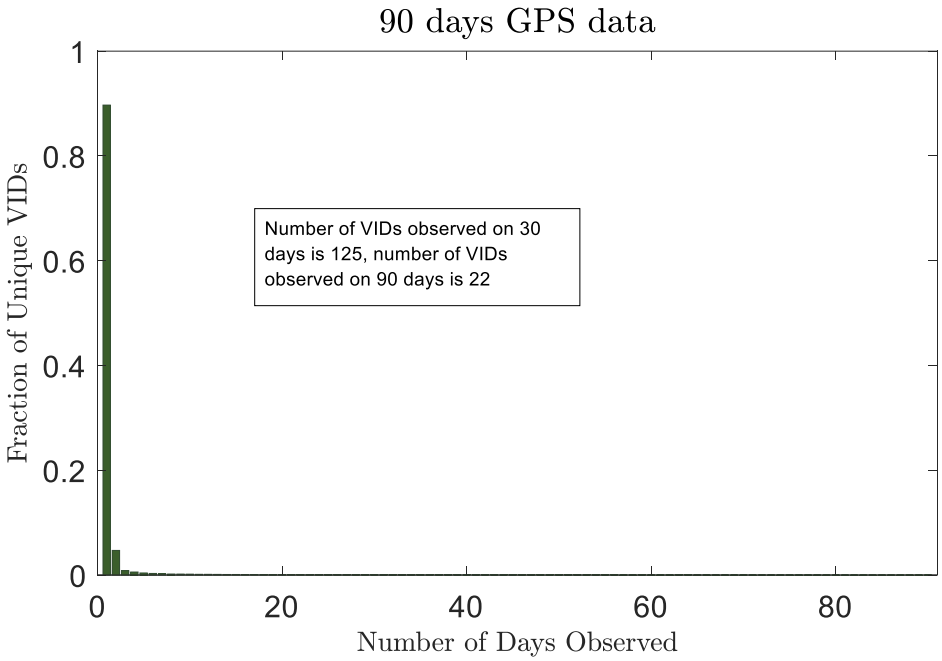
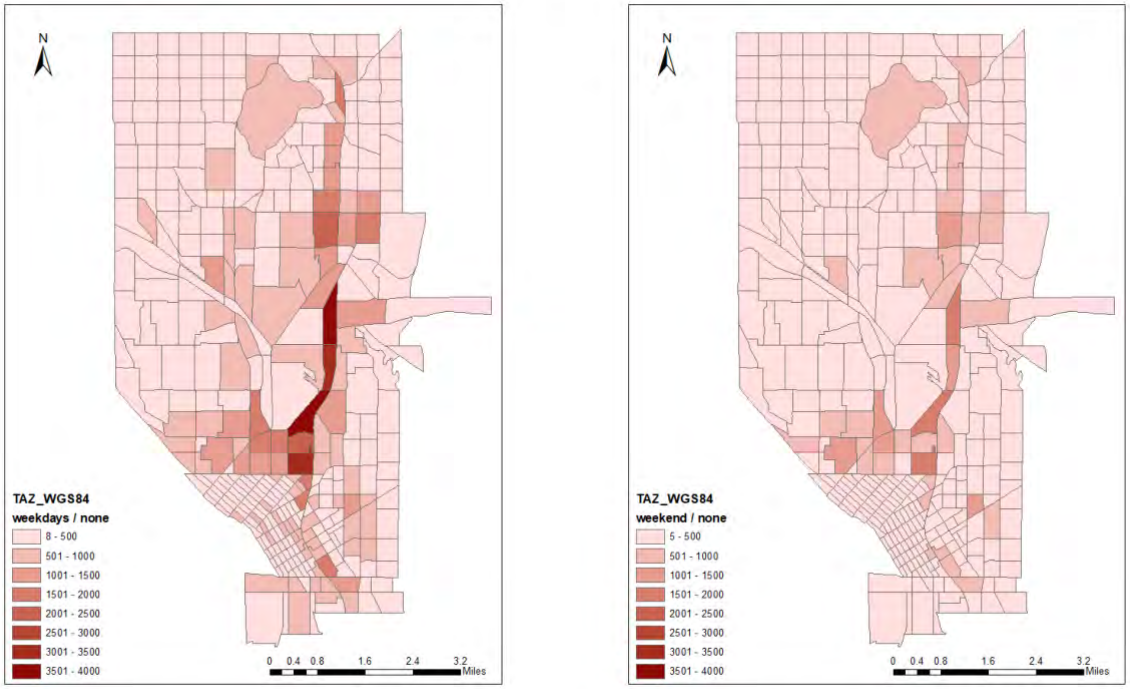


Figure 60. Graph. Number of days observed for unique VID.

5.1.5 Spatial Distribution of the Observations

The average number of GPS observations across different traffic analysis zones (TAZs) for weekdays and weekends are plotted in Figure 61. There are in total 360 TAZs in the study area. On average, weekdays have more observations than weekends. The TAZs that intersect with I-5 have a larger number of observations in both weekdays and weekends, further indicating that the GPS data correlates well with vehicular travel.



A. Weekdays B. Weekends

Figure 61. Graph. Spatial distribution of the observations.

5.2 First-order Properties

5.2.1 Activity Duration

Since GPS observations do not contain any semantic meanings such as the name or attributes of a place, algorithms need to be developed first to identify the activity locations from GPS data. A temporal-spatial clustering method (Ye et al. 2009) is used in this project to detect a “stay” (or stop) point and connect those stay points to construct trips. A stay point is a geographic area where a vehicle stays for at least a certain length of time. In the stay point detection process, if a GPS device consecutively produces a group of GPS points within a spatial region (less than 330 feet) for over a time threshold (300 seconds), this group of observations will be treated as a stay. Once stays are identified based on the temporal-spatial analysis, connecting two consecutive stays can produce a trip. Obviously, the trip timestamp and activity duration have certain inherent temporal relationships that are presented in Figure 62. For GPS data, the “stay time” between two consecutive trips is defined as the “activity duration.” More detailed description of the trip identification method can be found in Appendix B. Processing GPS Data.

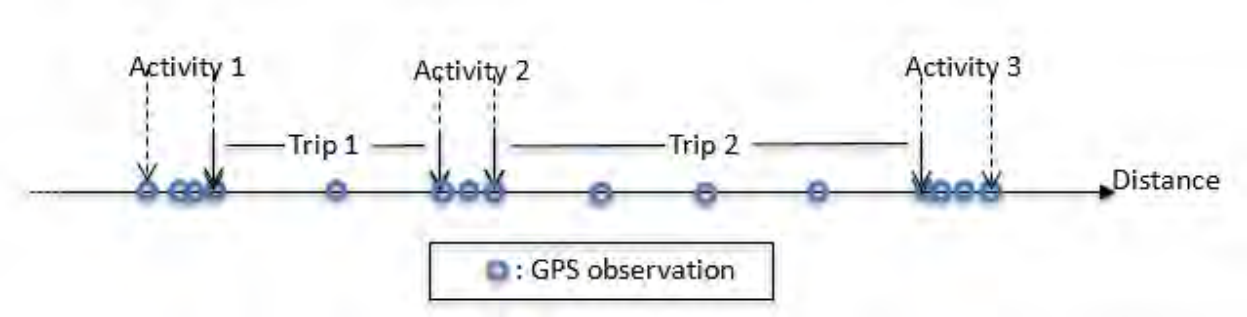


Figure 62. Illustration. Demonstration of the temporal-spatial relationships between trips and stays.

The activity durations derived from the GPS data vary from 5 minutes to 24 hours, because of the thresholds applied in the algorithm. For each day, around 80% of activity durations are less than 1 hour. Weekdays have more stays than weekends probably because on weekdays people have more work-related activities overall. Two examples for the number of stays in one day are shown in Figure 63: one for weekday, and another for weekend.

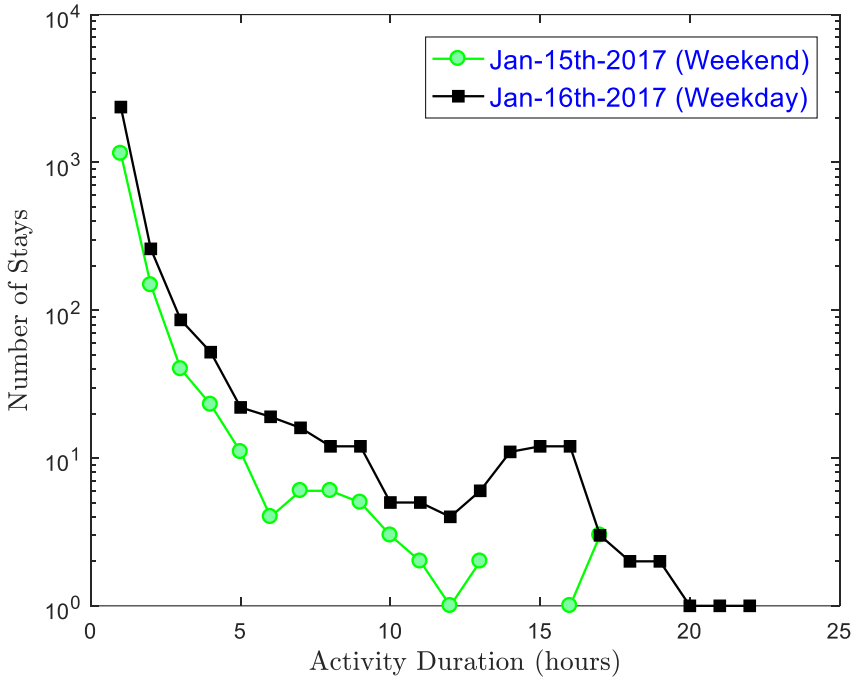


Figure 63. Graph. Distribution of activity duration.

For stays with a duration of less than 1 hour, the distribution of activity durations can be investigated. In Figure 64, each boxplot is for the percentage of stays falling into the specified one-minute interval (from 5 minutes to one hour). One could capture a monotone decrease trend for activity durations as the duration time increases, from 5 minutes to 30 minutes, and then stabilizes after 30 minutes. Data on weekdays and weekends show similar patterns, as do the 90 days of GPS data or monthly data, as shown in Figure 65 and Figure 66. As the length between

first and third quantile for each box is short, it implies that the activity duration patterns have small variations.

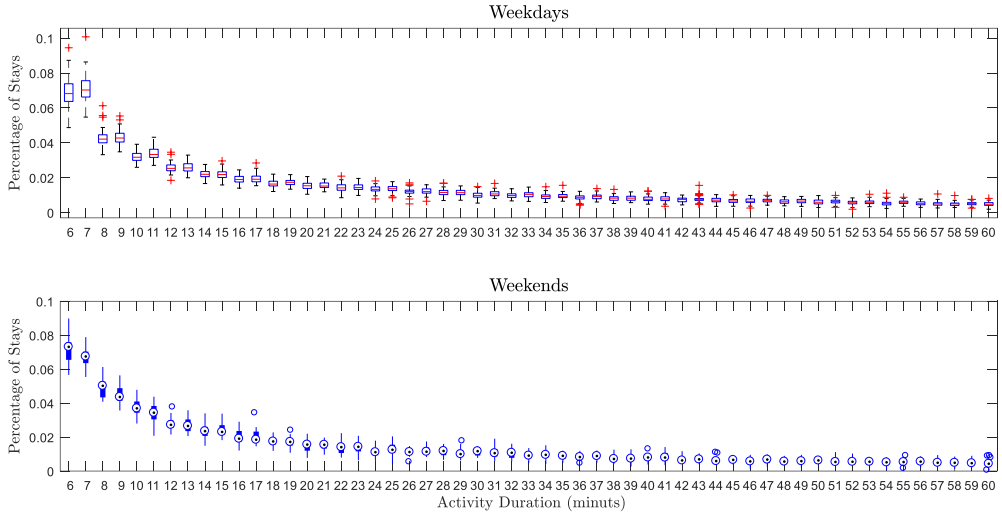


Figure 64. Graph. Activity duration for weekdays and weekends.

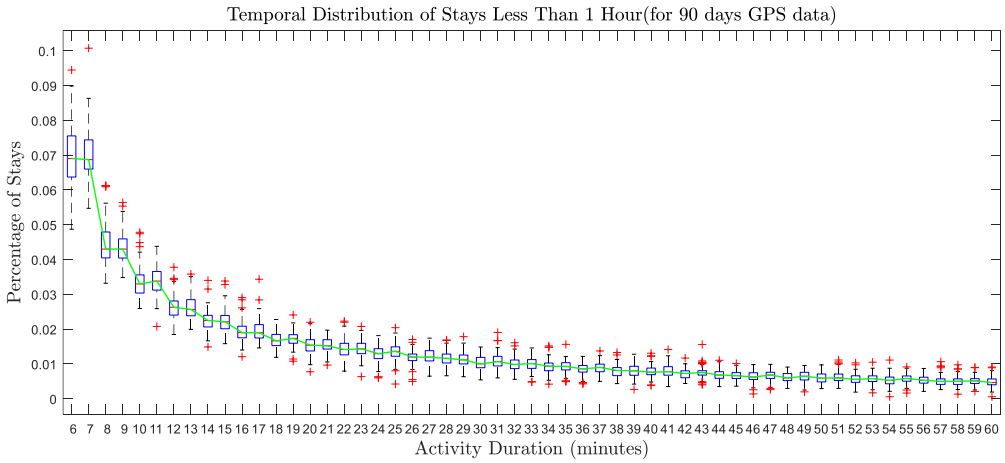


Figure 65. Graph. Activity duration for 90 days GPS data.

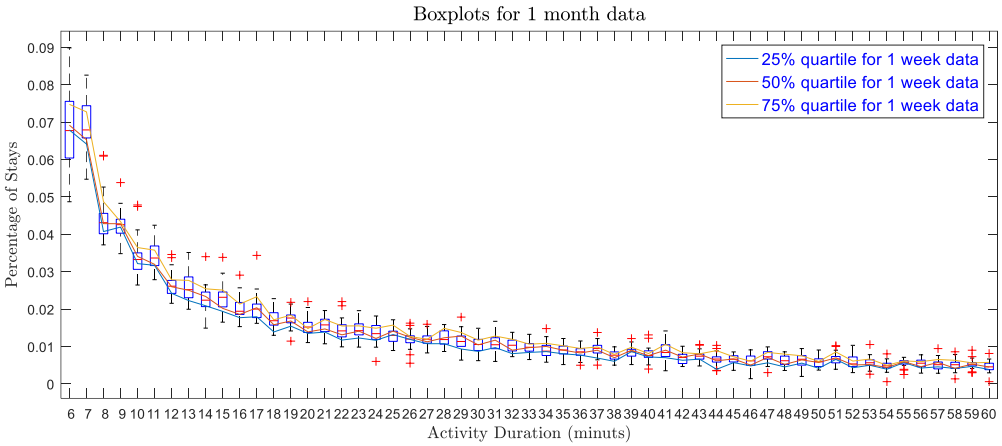


Figure 66. Graph. Activity duration for 1-month GPS data.

A stark difference is found while comparing the activity durations inferred from the GPS data with the household travel survey conducted by the Puget Sound Regional Council (PSRC, the regional metropolitan planning organization (MPO) in 2015. The survey data show a wide variation in activity duration, while according to the GPS results, most of the activity are rather short (see Figure 64). Several reasons exist for this difference in distribution. First, the GPS data is mostly for vehicular trips for a small area while the PSRC survey is for all trips in the entire Puget Sound Region. The GPS data cannot capture all types of trips, such as no-vehicular trips (e.g., walking) or those trips outside the study area. As discussed, the GPS data mainly capture short trips, and thus the results in Figure 67 are skewed heavily toward short trips. Secondly, the VID may have changed for the same vehicle. As a result, certain trips especially longer trips are particularly hard to capture by the GPS data.

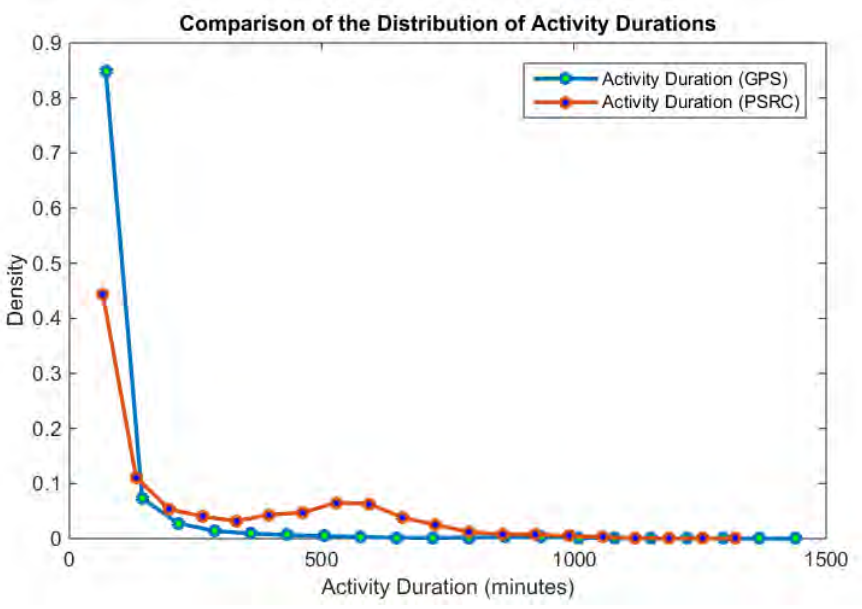


Figure 67. Graph. Comparison of distribution of activity durations.

5.2.2 Spatial Distribution of Zone-level Trips Origins (Generations) from a Zone

The observed numbers of trips originating from and destined to the 360 TAZs in the study area can be counted during weekdays and weekends. Figure 68 illustrates the number of trips originated from different TAZs on weekdays and weekends. Overall, the weekends have less number of trips when compared with weekdays. The TAZs that are closer to the downtown and University District originated a comparatively high number of trips during weekdays. While on weekends, areas such as Green Lake, which play a role as recreational locations, have a comparatively high number of trip originations.

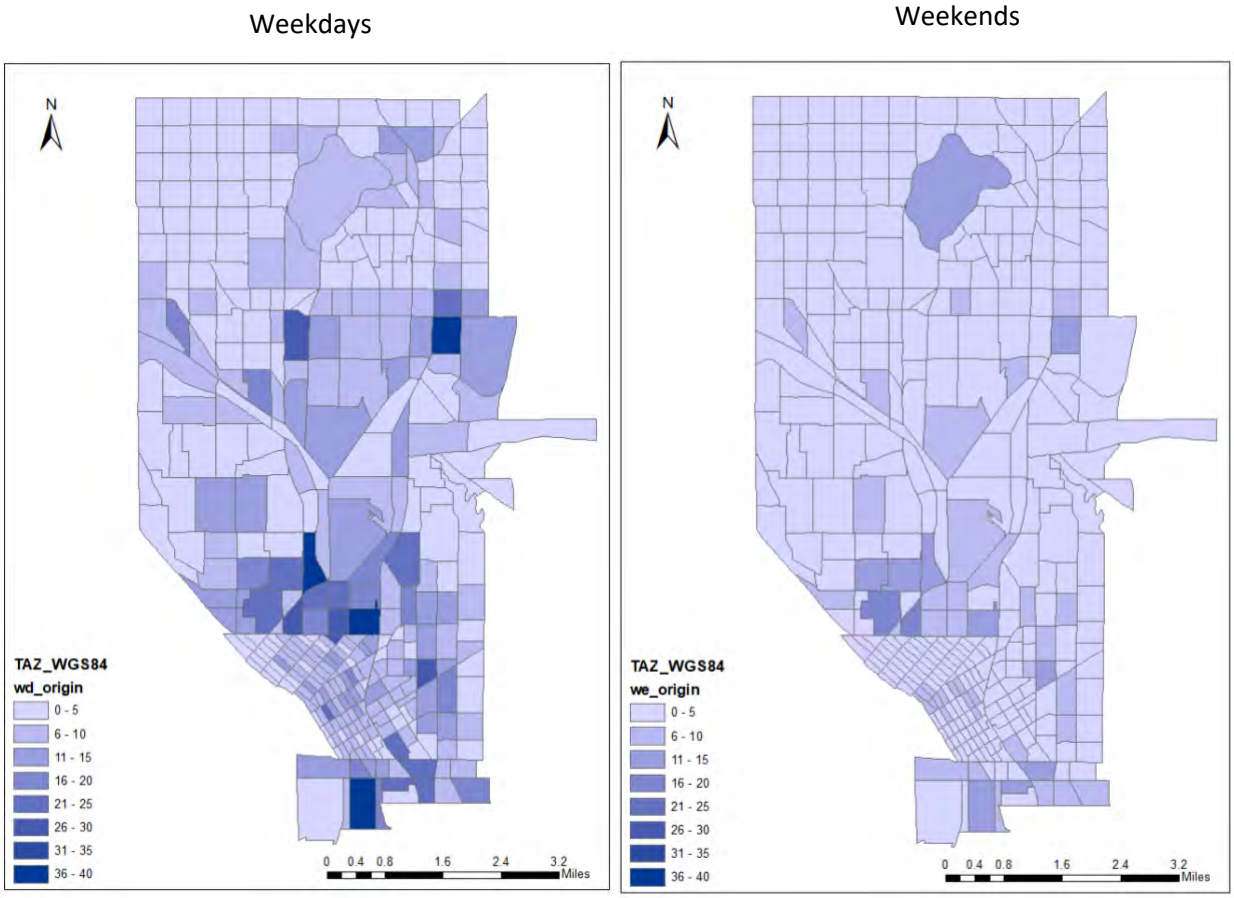


Figure 68. Graph. Number of trips originated from different TAZ on weekdays and weekends.

5.2.3 Correlation of Zone-level Trip Origins (Generations) With Population and PSRC Demands

Investigations are conducted next for the correlations between the observed number of trip origins from each TAZs with (i) the trip generation based on PSRC's travel demand model (base year 2014), and (ii) the population of the TAZ (base year 2014). The PSRC travel demand table used here is only for vehicular traffic. The observed number of trips originating from TAZs and PSRC travel demand have a correlation coefficient of 0.57 for weekdays and 0.54 for weekends. The

correlation with TAZ population is much weaker, with values of 0.17 and 0.14 respectively for weekdays and weekends. The almost nonexistent correlation between observed number of trips and population implies that the GPS data used here may not be considered as a representative sample of the population.

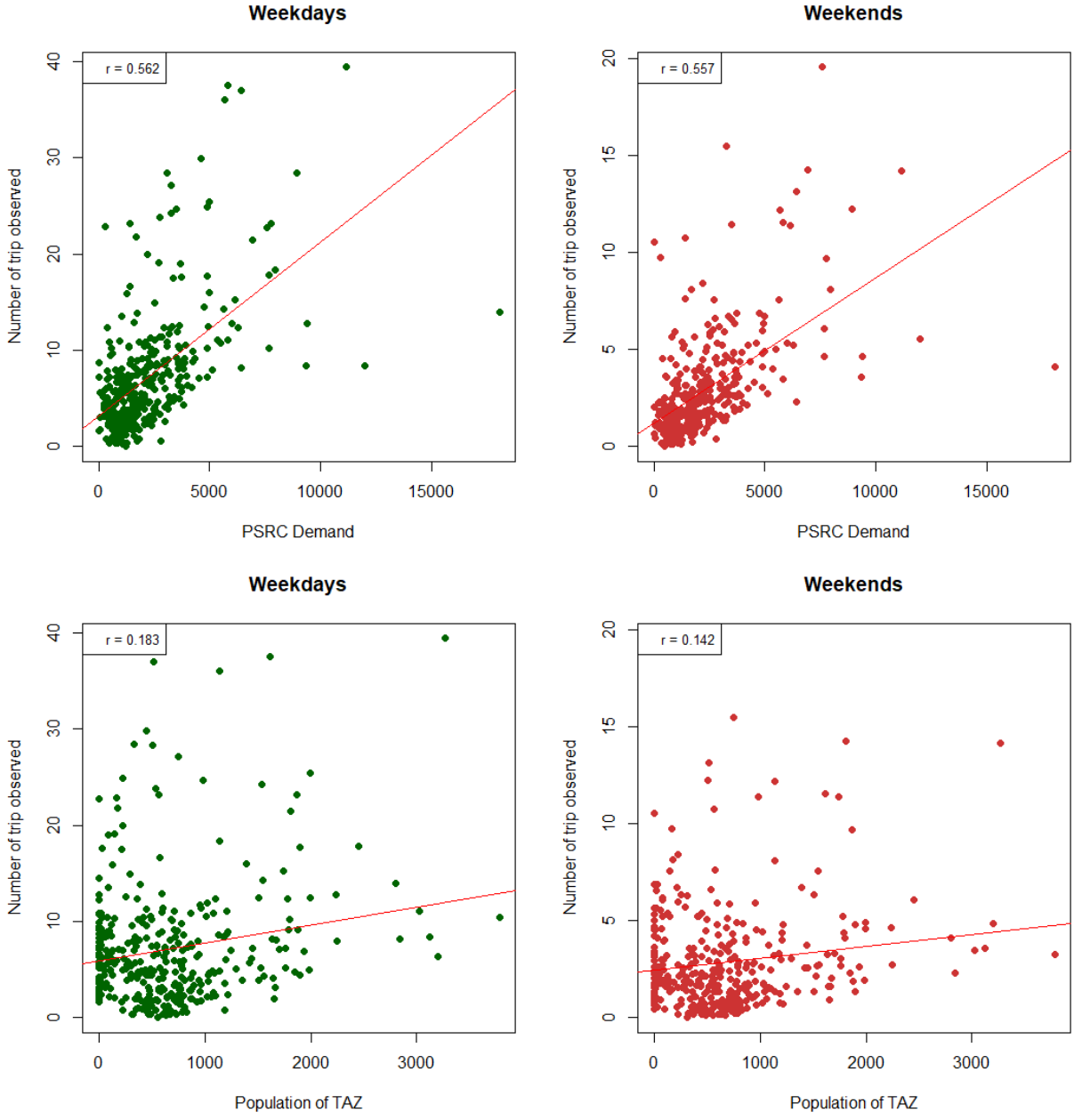


Figure 69. Graph. Correlation of TAZ-level trip origins with MPO results and population of TAZ.

5.2.4 Spatial Distribution of Zone-level Trips Destinations (Attractions) from a Zone

Investigations on trip destinations reveal similar spatial distribution among different TAZs for weekdays and weekends, compared with trips originating from TAZs, as shown in Figure 70. The TAZs near downtown and University District have a comparatively high number of trip destinations on weekdays, while recreational areas such as Green Lake have a relatively high number of trip destinations.

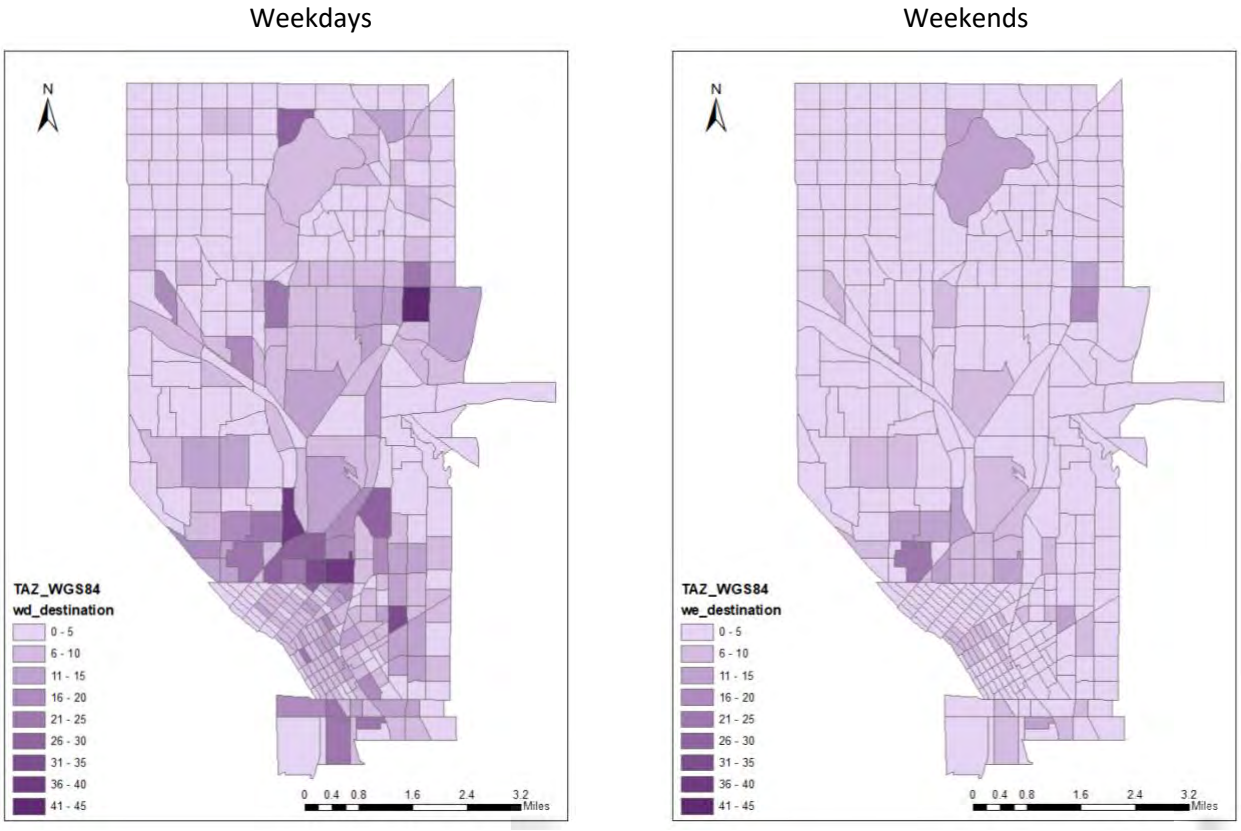


Figure 70. Graph. Number of trips destined to different TAZ on weekdays and weekends.

5.2.5 Correlation of Zone-level Trip Destinations (Attractions) With Population and PSRC Demands

The number of trips destined to each TAZ and the PSRC demands have a correlation coefficient of 0.56 for both weekdays and weekends, as shown in the scatterplots in Figure 71. The correlation with population gives similar results as trip originating from TAZs, with a correlation coefficient of 0.17 for weekdays and 0.14 for weekends.

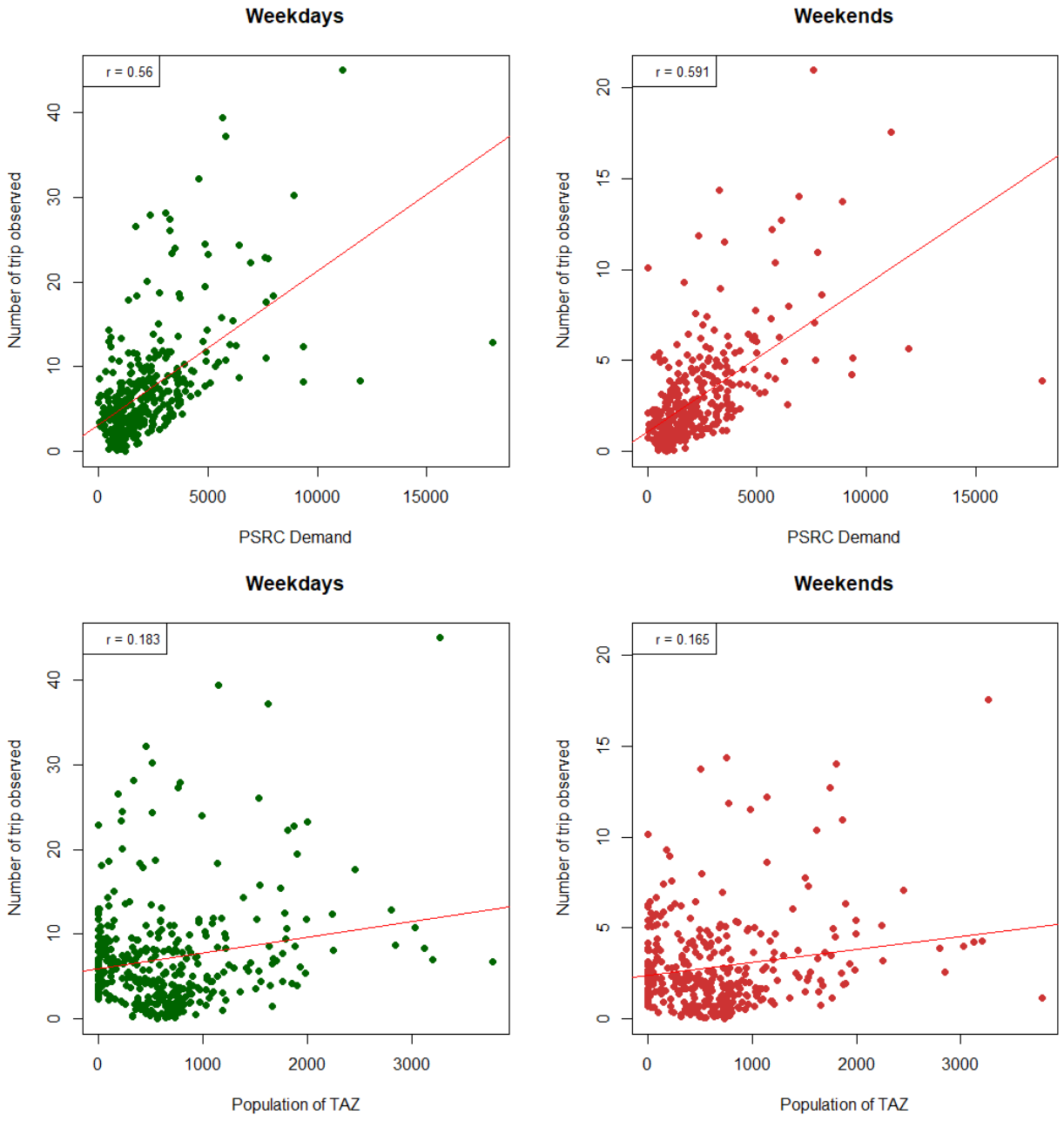


Figure 71. Graph. Correlation of TAZ-level trip destinations with MPO results and population of TAZ.

The correlation between trips derived from GPS data and regional travel demand are quite similar when compared with the results from mobile phone data. However, the GPS data show much weaker correlation with population data compared with mobile phone data. Such findings may have important implications when developing more advanced methods to fuse various data sources for more reliable and accurate OD estimation.

5.3 Second-order Properties

5.3.1 Distribution of Trip Rates

Figure 72 shows the histogram of trip rates of a VID within a day. Over 53% of VIDs produced only one trip per day, and over 94.22% of VIDs produced less than or equal to seven trips per day. Therefore, the trip rate of GPS data is relatively small. It can also be seen that the trip rate distribution has small variation, which indicates that the daily, weekly, and monthly patterns of trip rates from GPS data are similar. The low trip rates of a VID may result from multiple reasons. First, the inconsistency between VIDs and the actual vehicles (travelers) they represent, especially the fact that the same vehicle may be represented by multiple VIDs, may contribute to the low trip rates. Secondly, the observed trips of a VID may be only a part of its actual trips (i.e., there are unobserved trips of the VID), which is another reason for the low trip rates. The most important reason is probably that the study area is rather small and the travel activities of a vehicle (or VID) cannot be fully captured by the GPS data in this small area. As a result, the trip rates distribution is skewed to a smaller number of trips. This also implies that the trip rates distribution heavily depends on the study area. The distributions for small and large areas are expected to be different, as can be seen if Figure 72 (from GPS data) and Figure 33 (from mobile phone data) are compared.

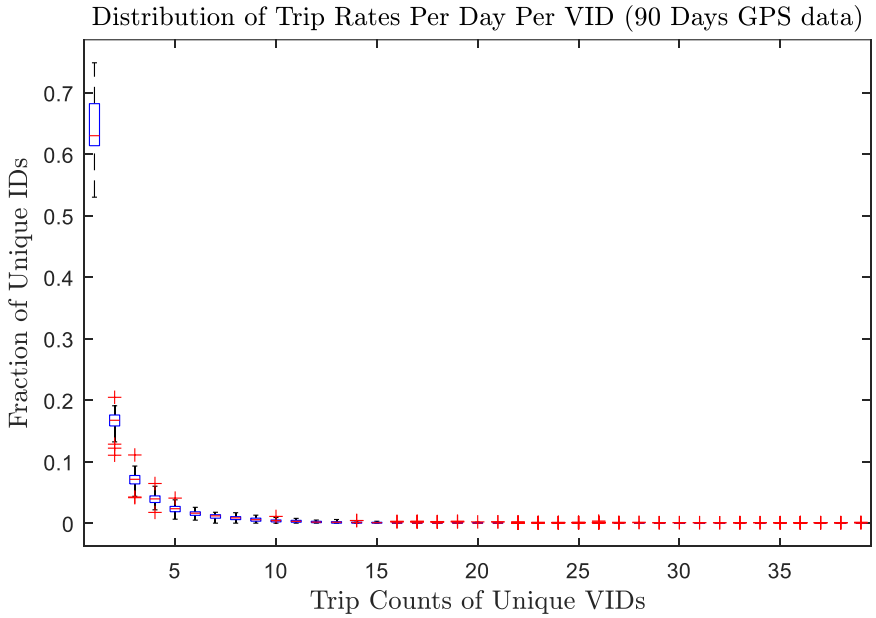


Figure 72. Graph. Distribution of trip rates.

Figure 73 illustrates a comparison of the distribution of trip rates with PSRC's household survey results. More trips were found in the survey dataset, very likely due to the reasons discussed above.

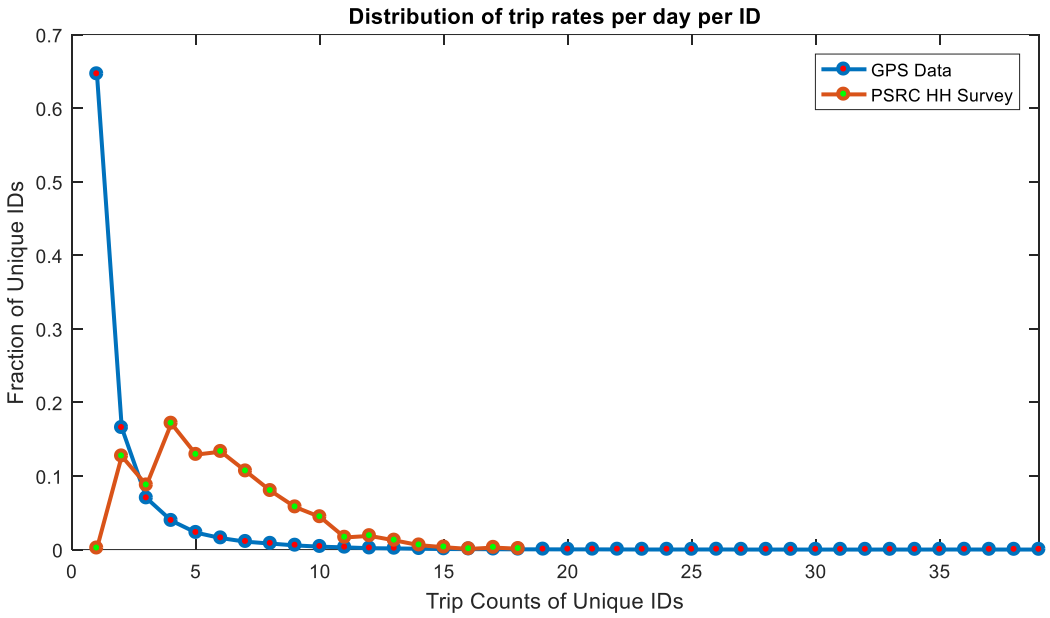


Figure 73. Graph. Comparison of distribution of trip rates with PSRC household survey results.

5.3.2 Distribution of Departure and Arrival Times

In order to investigate trip arrival time distribution on the time of the day, the number of trips ended (i.e., for arrivals) in each one-hour interval of a day (24 of them in total) is counted, and the fraction of those trips within the total number of trips of the day is calculated. The median of such fractions for the 90-day GPS data is then calculated. The results are shown in Figure 74-A and B, for weekdays and weekends, respectively. As over 86.38% of the trip durations are less than 15 minutes, the departure time distributions show similar patterns as the arrival time distributions, shown in Figure 74-C and D. The focus here is only on the arrival time distributions.

The trip arrival time distributions on weekdays and weekends have their own peak periods and off-peak periods. Some dissimilarities exist in at least two aspects. First, the trip arrival fraction of weekdays increases dramatically from 5:00 to 8:00 and keeps at a relatively high level until 16:00, which reflects intense activity level during the daytime on weekdays. On weekends, the fraction of trip arrivals gradually rises starting at 6:00 and achieves its peak at around 13:00. From 6:00 to 8:00, the variations of fraction of trip arrival on weekends are large, which indicates that there are morning peaks on weekends for some weeks, but not for other weeks. The median of the fractions for weekdays maintains a high level of 8 hours during the daytime, while the median of the fractions for the weekends shows a unimodal distribution and reaches its maximum value in the late noon. Second, during the daytime (7:00-17:00), the fraction of arrivals for weekdays is always greater than that of the weekends, and the two fractions have the maximum difference during morning the peak hours. In the night time (17:00-7:00), however, the fraction of arrivals of weekends is slightly larger than that of weekdays. This difference continues to show travelers' various travel patterns on weekdays and weekends. For example, trips on weekdays are mostly home- and work-based trips that happen mostly during the daytime. On weekends, however, people's travel purposes are more likely to be recreational activities for leisure; most trips do not

follow the working-hour schedule. Similar to what has been discussed in Section 5.1.3, this finding also implies that, with proper processing (e.g., trip ends identification), GPS data may be used to produce reasonable trip arrival/departure patterns. This is because the GPS data were collected mainly from vehicular travel related devices/apps, although further validation with some "ground-truth" data (such as loop detector data or survey data) is still needed to confirm this finding. The comparison with the PSRC survey results are provided later in this section.

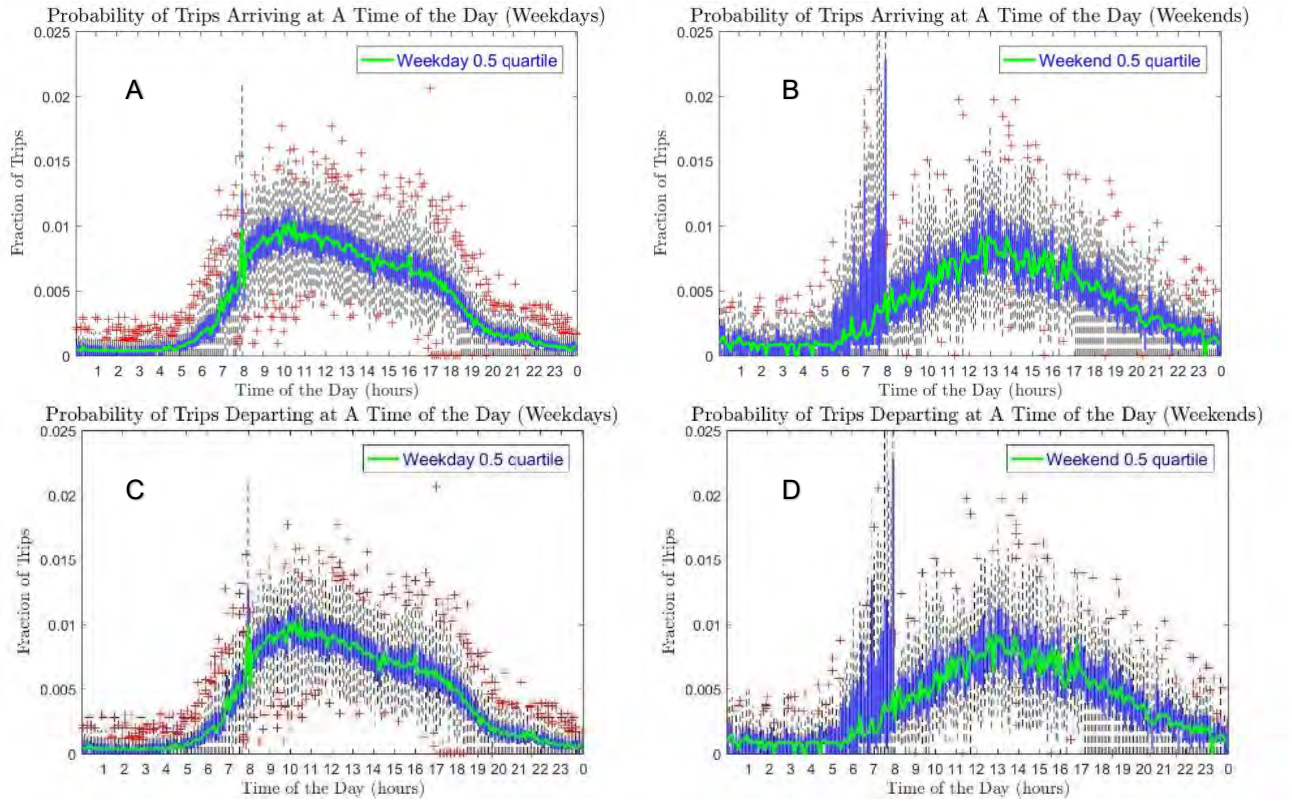


Figure 74. Graph. Distribution of departure and arrival times.

In addition to comparing different travel patterns between weekdays and weekends, it is also interesting to look into how those patterns change as GPS data with longer durations are considered. In Figure 75-A and B, it is found that for weekdays and weekends, when 1-week, 5-week, or 90-day GPS are considered, the fraction of trip arrivals of each interval varies from scenario to scenario. For weekdays, those variations are not significant except for early noon. Compared with weekday data from 90 days, the fractions of trip arrivals calculated from smaller datasets (one week) and larger datasets (3-week and 90-day datasets) have less than 1% difference. On the other hand, for weekends, in addition to the large difference in the early morning (6:00-9:00), significant differences exist from late afternoon (13:00) to early evening (20:00). In conclusion, to investigate the trip arrival time distributions, GPS data from one or fewer weeks are less reliable, especially for the distributions of weekends. If possible, one or even a few months of GPS data are preferred to produce more reliable and stable results. Since most trips have a travel time of less than 15 minutes, similar patterns (and similar conclusions) present in the

departure time distribution compared with the distributions of the arrival times, as shown in Figure 75-C and D.

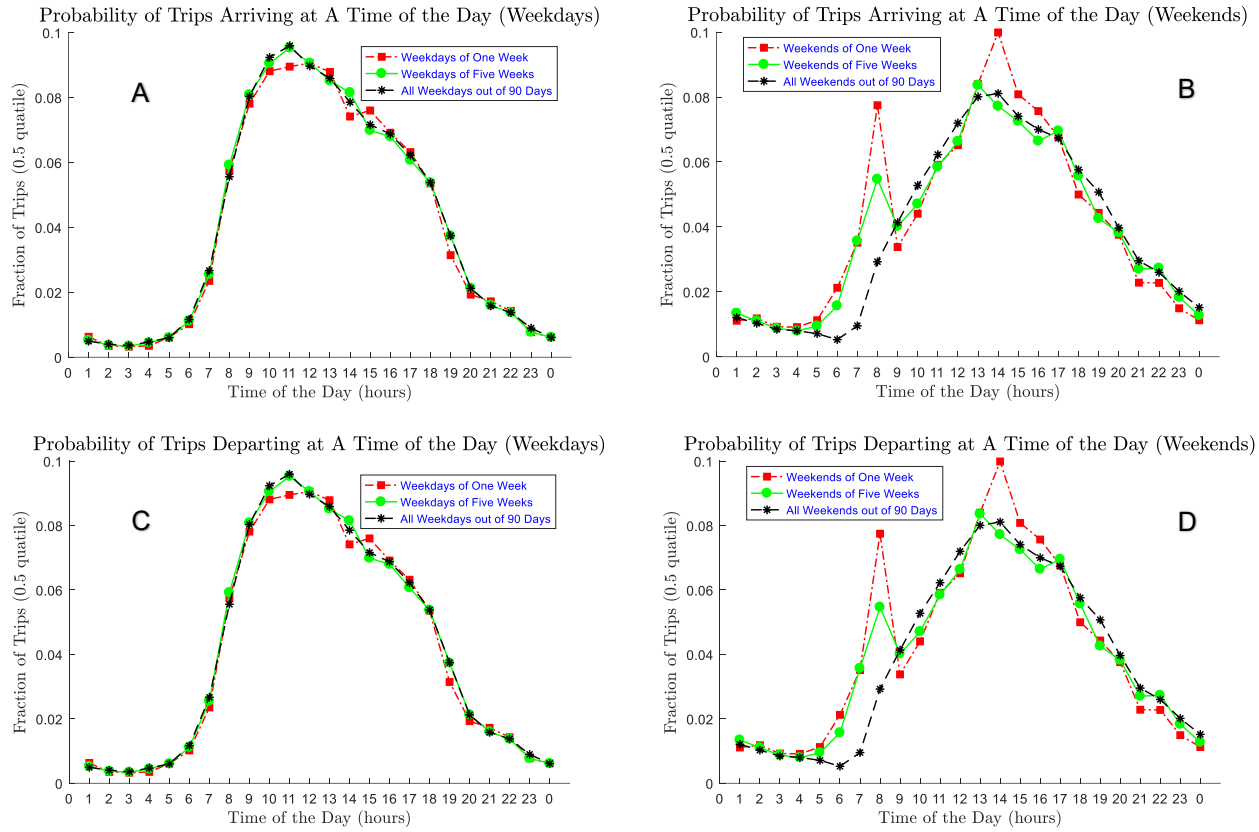


Figure 75. Graph. Distribution of departure and arrival times for different data size.

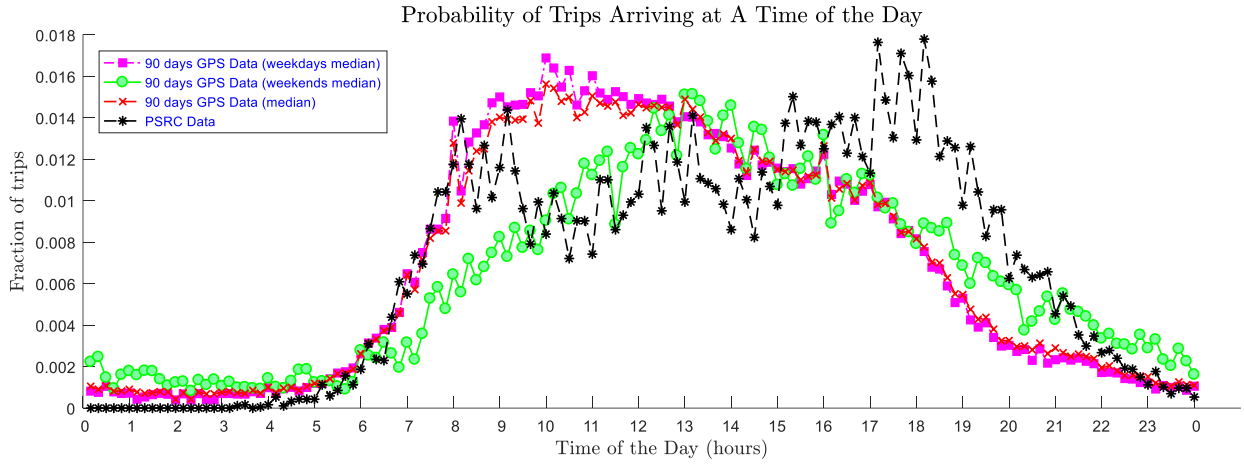


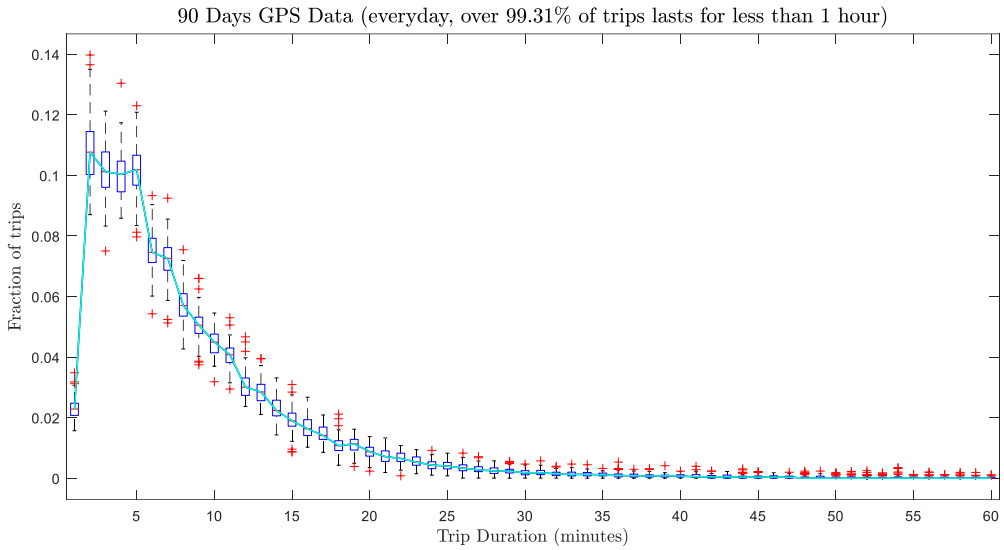
Figure 76. Graph. Comparison of trip arrival times with PSRC household survey data.

Figure 76 further illustrates the distribution of arrival times from the GPS dataset along with that from PSRC household survey. Unlike the GPS dataset, two distinct peaks (morning and afternoon) can be found in the survey results (shown in black). The discrepancy with the PSRC

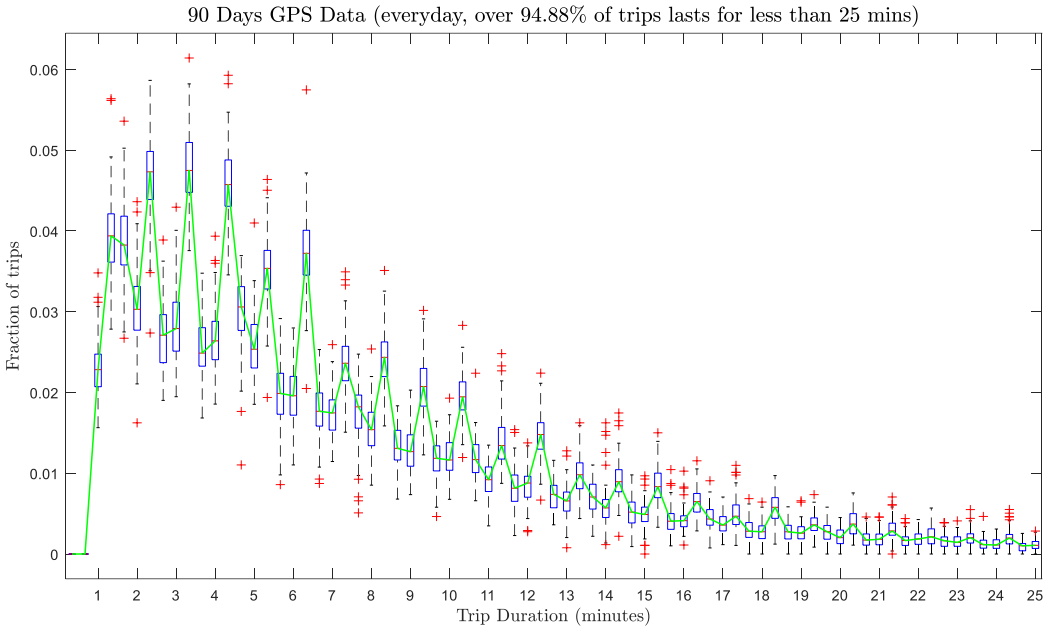
survey data in terms of the arrival patterns is probably due to the same reasons as discussed above for the trip rates in Figure 73, especially the fact that the study area is relatively small.

5.3.3 Distribution of Trip Times

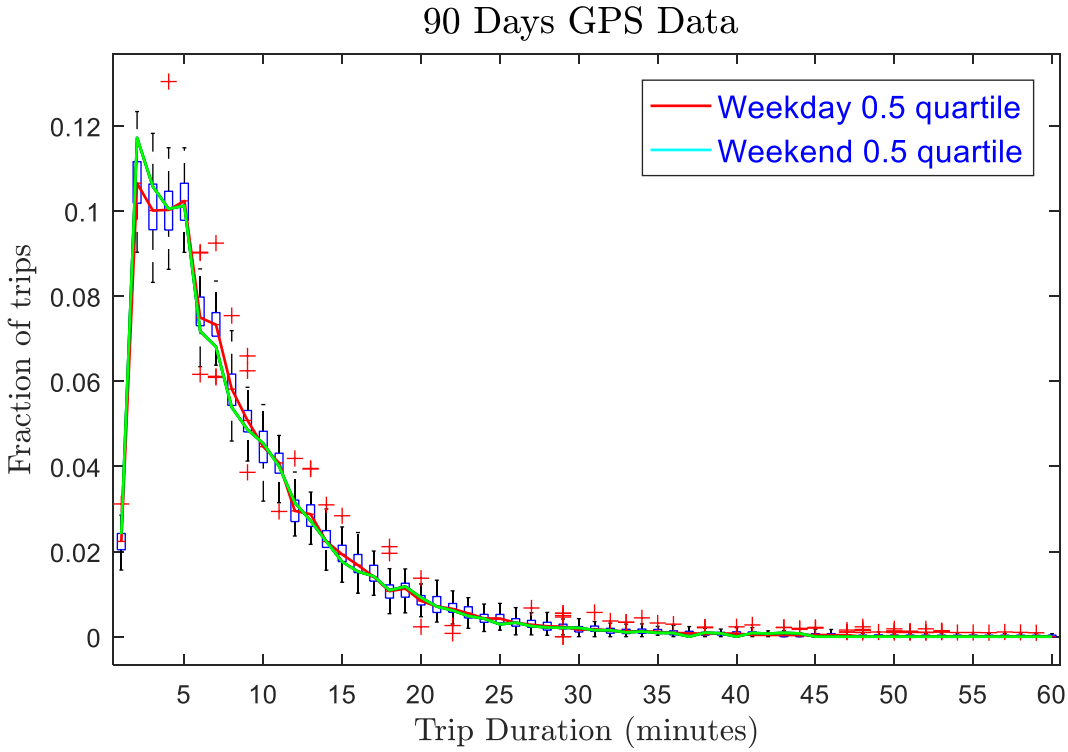
As shown in Section 5.2.1, out of all the trips identified from the GPS data, over 99.31% lasted for less than 1 hour. From Figure 77-A, the trips whose durations are in the interval [2, 10] minutes take the largest portion. This may be explained by the fact that the GPS data study area is relatively small and includes only part of the Seattle downtown, which is typical for frequent and short trips. It can be seen that the variation of trip durations is small: the difference of 0.25 quartile and 0.75 quartile is small for most boxplots. The weekday 0.5 quartile and weekend 0.5 quartile show a similar pattern as the overall distribution pattern, as shown in Figure 77-C. In addition, daily trip duration distributions share similar patterns throughout the 90 days, and 94.88% of the trips from the 90-day data lasted for less than 25 minutes. Figure 77-B shows the trip time distribution for trips less than 25 minutes. In Figure 77-B, each boxplot stands for the fraction of trips belonging to a time interval of 20 seconds. As the time interval of each boxplot is small, the median value in Figure 77-B has more fluctuations than in Figure 77-A.



A. Less than one hour.



B. Less than 25 minutes.



C. Median trip fraction on weekdays and weekends.

Figure 77. Graph. Distribution of trip durations for different data size.

Figure 78 illustrates a comparison of trip durations between the trips extracted from the GPS dataset and the PSRC travel survey data. Longer trips can be found in the travel survey dataset while compared to the travel survey results. The main reason for this is probably the small study area in GPS dataset, which greatly reduces the prevalence of trips longer than 20 minutes.

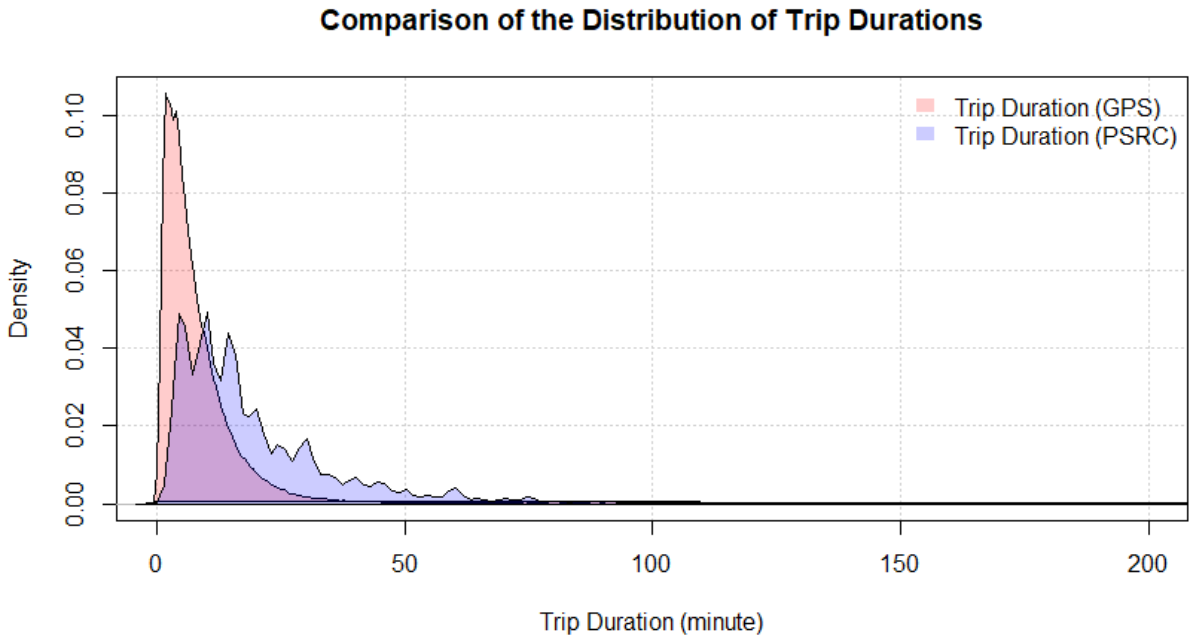


Figure 78. Graph. Comparison of distribution of trip duration with PSRC household survey data.

5.3.4 Percentage of Zero cells

The properties related to zero cells are investigated, which are defined as the OD pairs that do not have any observed trips from the GPS data. These are important features of the data since the applied upscaling method will result in zero trips in the OD table for those OD pairs. Figure 79 first illustrates the change in the percentage of zero cells with the number of days of GPS data that are used to calculate the OD table. The percentage of “zero-cells” monotonically decreases with the increase of the observation period. The percentage reaches the lowest value of around 64% when data from the entire study period (90 days) are used. It means that 64% of all OD pairs still do not have any observed trips between them considering the entire study period of 90 days. Notice that the PSRC travel demand table only contains a handful of zero cells (less than 10), indicating that the GPS data only captures a small friction of actual travel.

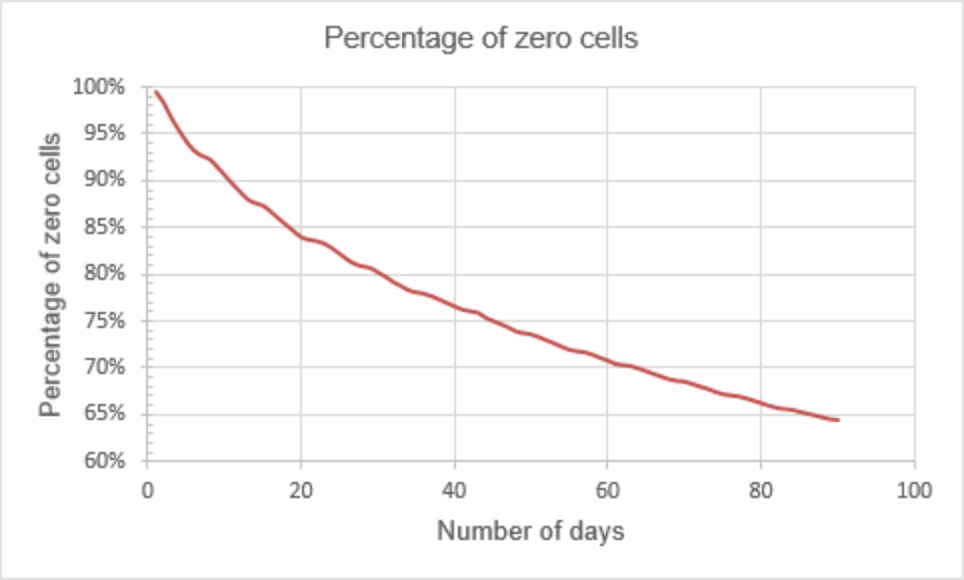


Figure 79. Graph. Change in the percentage of cells with zero trips with increase of the study period.

Investigations are conducted next regarding the spatial distribution of the TAZs that have zero trips originated from and destined to. Considering each TAZ as origins, the total number of destination TAZs with zero trips from the origin is first calculated. Similarly, considering each TAZ as destination, the total number of origin TAZs with zero trips to the destination is calculated. These two numbers are added and then divided by the sum of total number of origins and destination TAZs (from or to the subject TAZ) to obtain the percentage of the sum of zero trips for each TAZ. In mathematical terms,

Percentage of the sum of zero trips for TAZ$(i) = \frac{N_{O,i} + N_{D,i}}{2N} \times 100\%$

Where,

$N_{O,i}$ = Number of TAZs that have zero trips originated from TAZ (i)

$N_{D,i}$ = Number of TAZs that have zero trips destined to TAZ (i)

N = Total number of TAZ in the study area (360)

Figure 80 illustrates the percentage of the sum of zero trips (both origin and destination) for TAZs in the study area for the entire study period of 90 days. Most zones with a higher percentage of zero cells lie in the borders of the study area. The reason for this pattern in the border TAZs might be the partial inclusion of these TAZs in the study area.

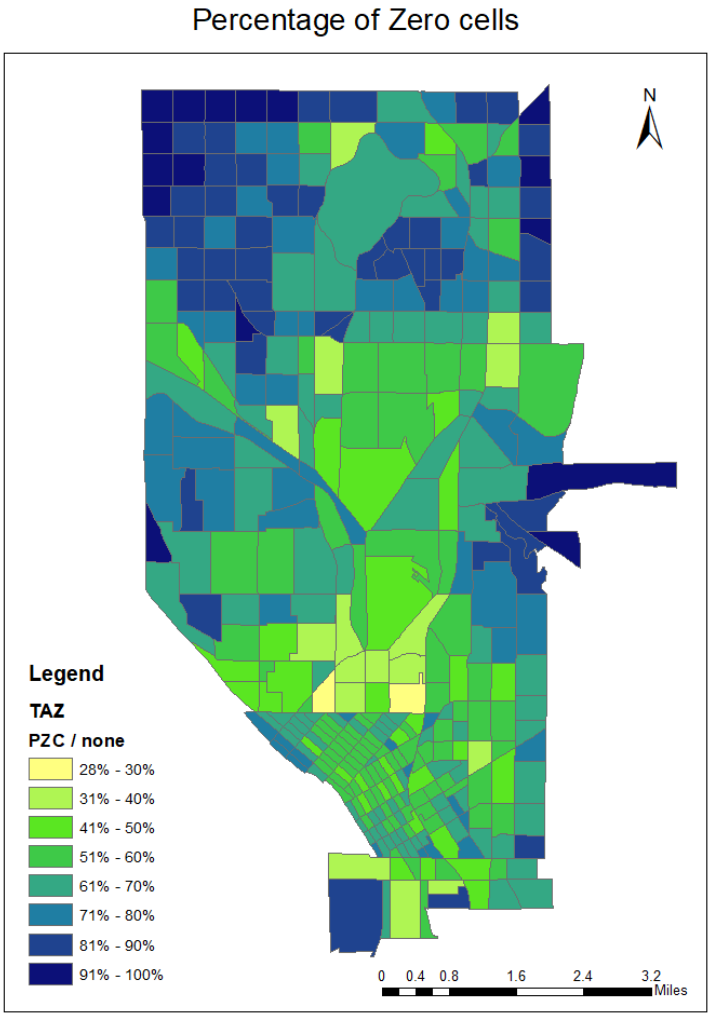


Figure 80. Graph. Spatial distributions of TAZs with zero trips between OD pairs.

5.3.5 Spatial Distribution of Zero Cells

Similar patterns can be observed if the percentage of cells with no trips generated from or attracted to a TAZ is shown separately; see Figure 81. Most TAZs in the borders of the study area have a high percentage of zero cells while considering trips generated from or trip attracted to the zones.

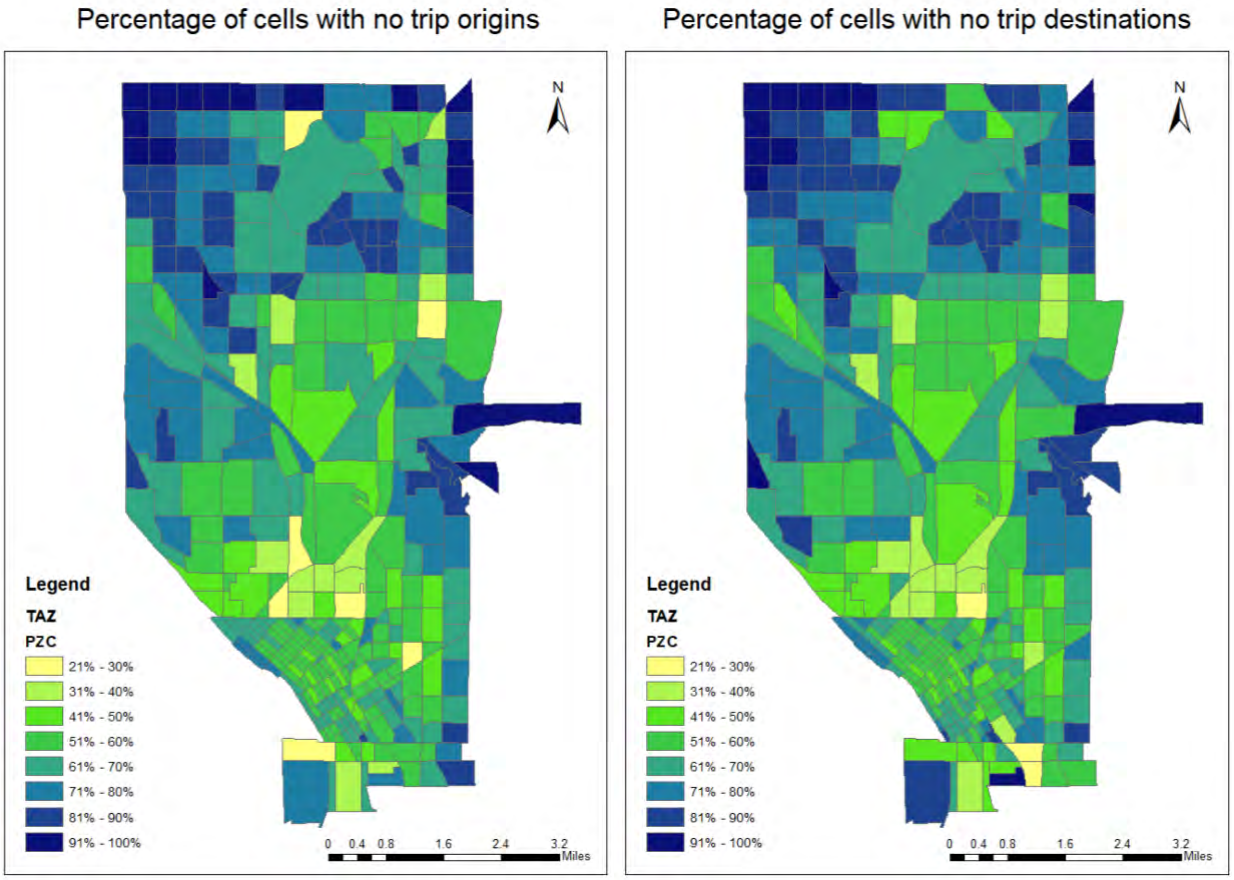


Figure 81. Graph. Spatial Distribution of TAZs with A) Percentage of cells with no trip origins. B) Percentage of cells with no trip destinations.

5.3.6 OD Collinearity

The above second-order properties focus on observed trips, which are used to calculate the origin-destination (OD) trip table for the study area. For this, the upscaling step is needed. In this research, the method developed by Wang et al. (2012) is used to upscale the observed trips. The pseudocode for this method is given in Appendix C. Pseudocode for OD Estimation from GPS Data.

To upscale the observed trips into (estimated) total trips originated from and destined to each TAZ, several parameters need to be applied according to the upscaling method proposed in Wang et al. (2012). They applied the method on large-scale mobile phone data available in the form of mobile phone billing records. One of the key tasks for applying the method is to infer home locations from the collected data. It is discovered in this project that finding the home location of a user using GPS data is difficult. This could be due to the large and irregular GPS data sampling intervals or the removal of the home-related data due to privacy protection reasons. Another important reason is due to the small size of the study area such that only a small fraction of vehicles appeared for multiple days, which however is crucial for the identification of home locations.

To estimate ODs from GPS data, this research assumes that the identified trip ends (discussed above) represents home or work places, so that the existing method (e.g., Wang et al., 2012) can be applied. This study also makes some modifications regarding the input parameters needed to upscale the observed trips into OD tables include the following:

1. The ratio of the population and the number of unique VIDs for each zone (M), which aims to estimate the (reciprocal of the) market penetration of the GPS devices in the GPS dataset for a zone.
2. Vehicle usage ratio (VUR) of each zone.
3. Total daily trip production for the entire population (W).

Among these parameters, the first two parameters are TAZ based and the last parameter is for the entire population. Since the TAZ-based population and VID data are available, the first parameter can be readily computed for each TAZ. The results from upscaling the trips using different $M's$ for different TAZs are compared with the result if only a common M is used for the entire study area. The common *M* is calculated as the ratio of the total population of all TAZs in the study area and the total number of observed unique VIDs.

The second parameter VUR, represents the percentage of the GPS observations that are from vehicles. The VURs for different TAZs can be calculated using the following equation (Wang et al., 2012):

$$VUR(i) = P_{car\ drive\ alone}(i) + \frac{P_{carpool}(i)}{S}$$

Figure 82. Equation. Definition of *VUR*

where, $P_{car\ drive\ alone}$ = probabilities that travelers in zone i drive alone
$P_{carpool}(i)$= probabilities that travelers in zone i share a car
S = Average carpool size

Ideally, VUR should be TAZ based. In practice, TAZ-based VUR is difficult, if not impossible, to obtain. In this project, an average VUR (0.346) is estimated that can be applied to all TAZs in the study area. Details of calculating VUR is given in Appendix C. Pseudocode for OD Estimation from GPS Data. Using *M*, VUR, and *W*, one can upscale the observed number of trips originated from and destined to each TAZ.

Figure 83 first illustrates a comparison of the results obtained using a separate M for each TAZ and a common M for the study area. Using separate *M*'s for different TAZs results in different upscaling factors for each TAZ, which means the estimated (total) OD trips do not maintain high correlation with the observed trips. In the case of using a common *M*, the upscaled OD remains highly correlated with the observations for all TAZs. In this study, since the GPS data cannot present the zone-level population well (see Figure 69), zone-level M's calculated above may not effectively represent the GPS device penetration at the zone level. The common *M*, by aggregating population and observations from all zones, is expected to better represent the area-level GPS device penetration. Therefore, the common *M* is applied to the entire study area for

OD estimation in this report. In any case, *how* the M is calculated does contribute to the differences between the OD table derived from GPS data and the PSRC trip table.

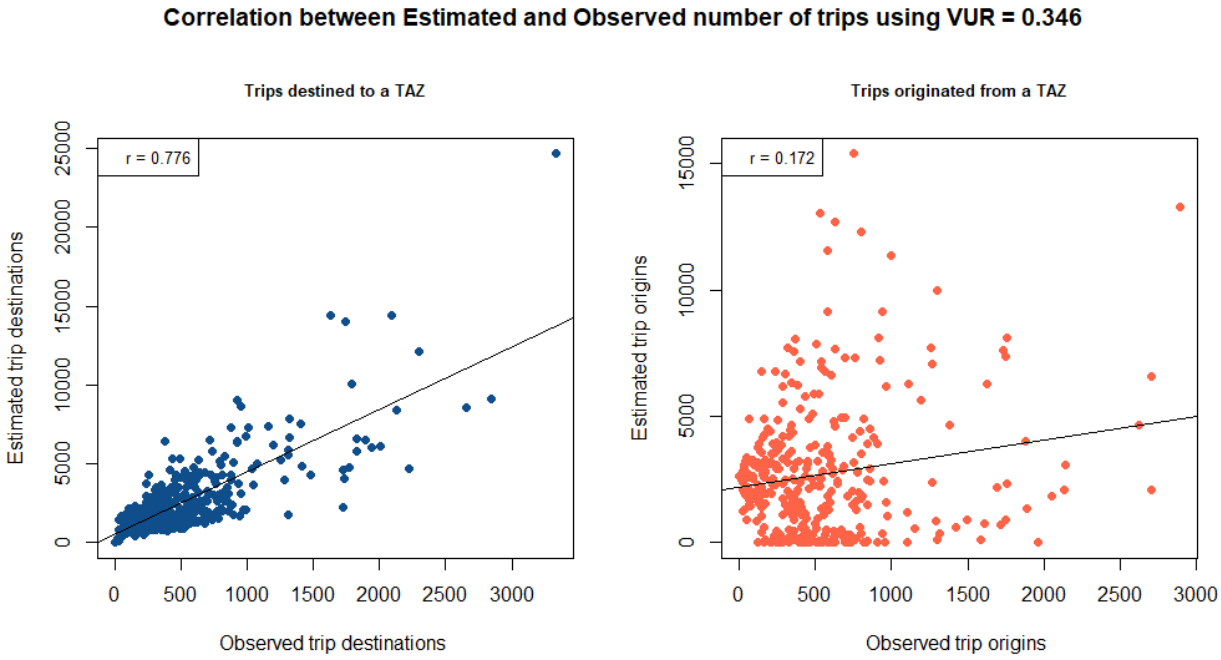


Figure 83. Graph. Correlation using separate M for each TAZ.

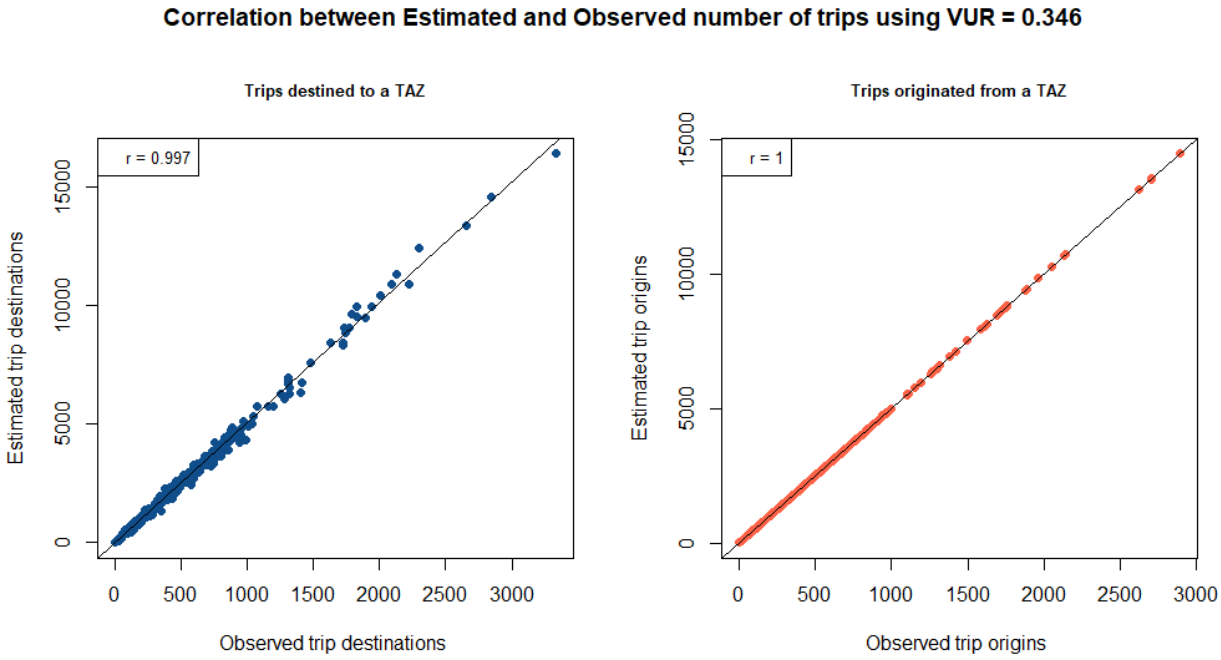


Figure 84. Graph. Correlation using common M for entire study area.

5.3.7 Comparison of the estimated OD table with the MPO OD Table

The estimated total OD trips originated from and destined to different TAZ have a correlation coefficient of 0.58 and 0.59, respectively, with the corresponding PSRC OD demands, as shown in Figure 85. The effect of zero cells between OD pairs in the estimated OD table from GPS data becomes evident by this comparison. The cells with no trips cannot be upscaled with the upscaling method used in this project. This is one important reason for the relatively small coefficients. When comparing upscaled OD trips from the GPS data to the PSRC demand, the correlation is even smaller at 0.347 (see Figure 87).

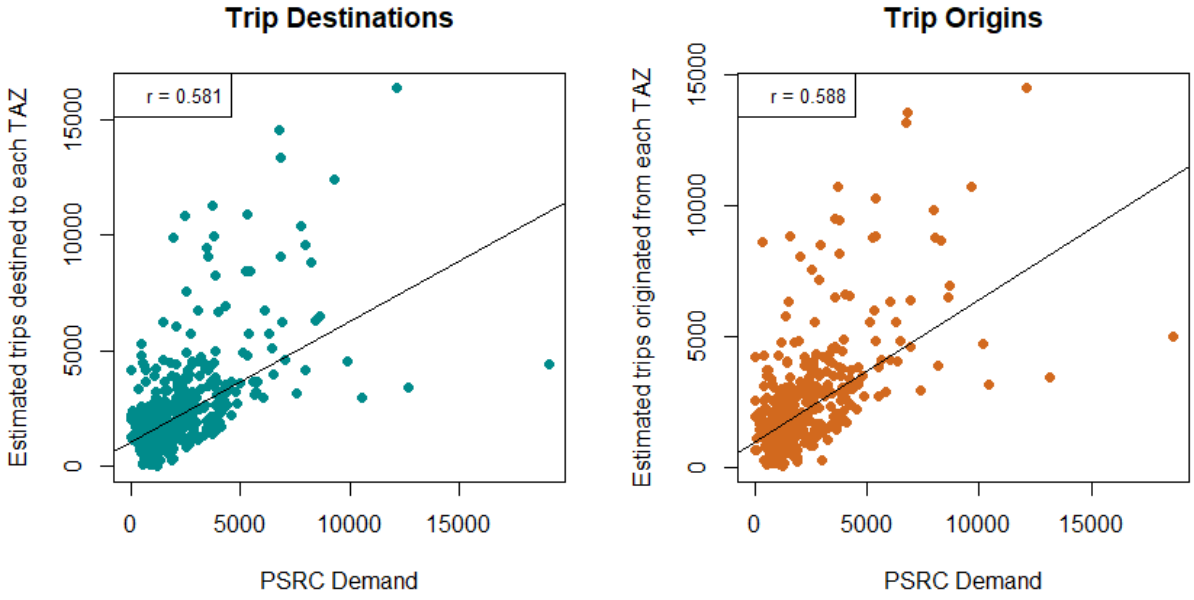


Figure 85. Graph. Correlation of estimated trip generation and trip attraction with corresponding PSRC demands.

The comparison between the estimated OD and the PSRC results is further illustrated in Figure 86. The lack of observed trips in the zones near the boundary results in lower number of trip originated from and destined to those zones. On the contrary, some zones (like Green Lake, Lower Queen Anne, and South Lake Union) exhibit higher numbers of trip originated from and destined to those zones while compared to PSRC demand. The reason for such differences may be due to the comparatively high number of observed trips in those zones and how the parameters are calculated in the upscaling method.

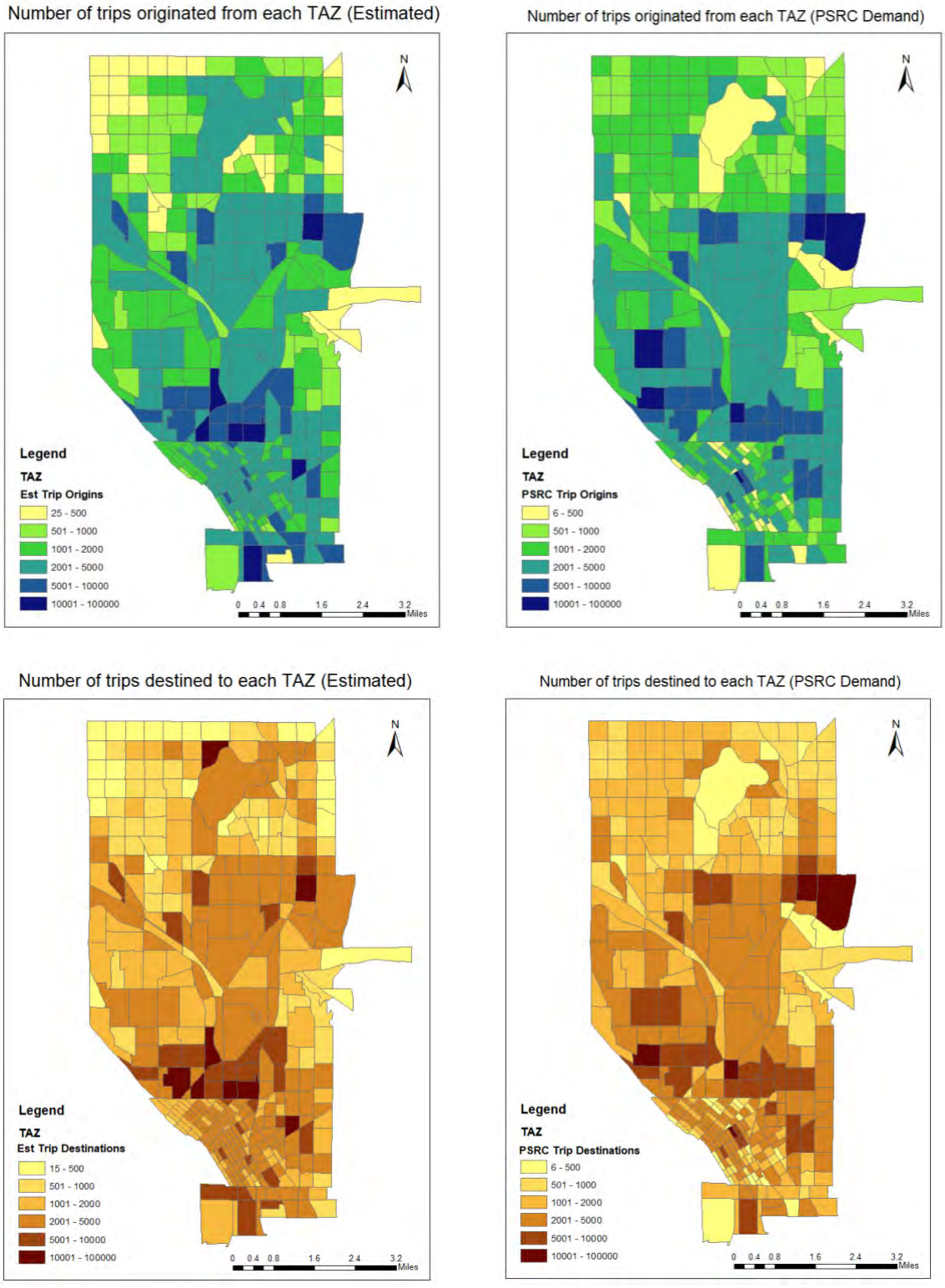


Figure 86. Graph. Comparison of estimated number of trips originated from and destined to each TAZ with PSRC demand.

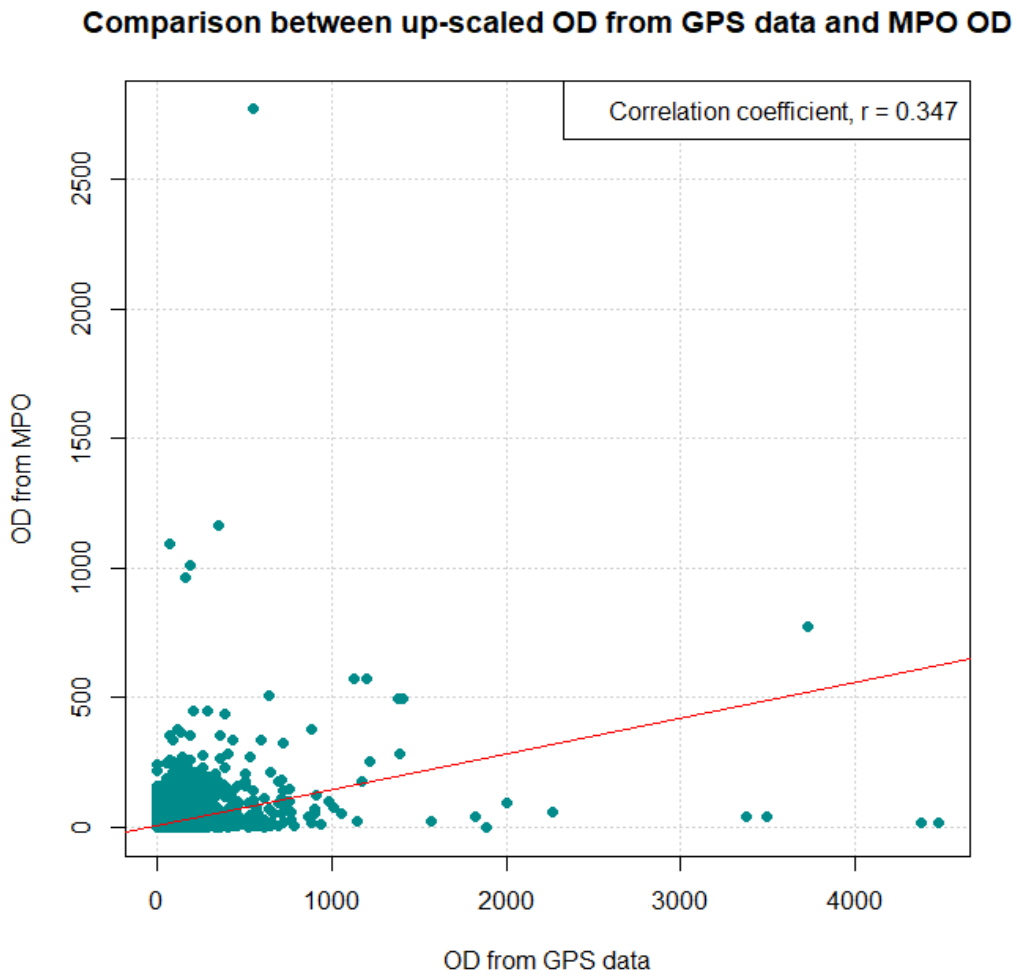


Figure 87. Graph. Comparison of up-scaled OD trips compared to PSRC demand.

5.3.8 OD Sensitivity Analysis

Besides the *M* parameter, zone-level vehicle usage ratio (VUR) is also hard to obtain. A sensitivity analysis is conducted here to see if one (randomly) varies VURs for different TAZs, how the OD results may change. For the analysis in this subsection, the common *M* parameter is used for the entire study area. In particular, if all VIDs from each TAZ are considered to be vehicle users who drive along (i.e., $VUR = 1$), then the estimated OD after upscaling the observed trips would be perfectly collinear to the observed trips as seen in Figure 88.

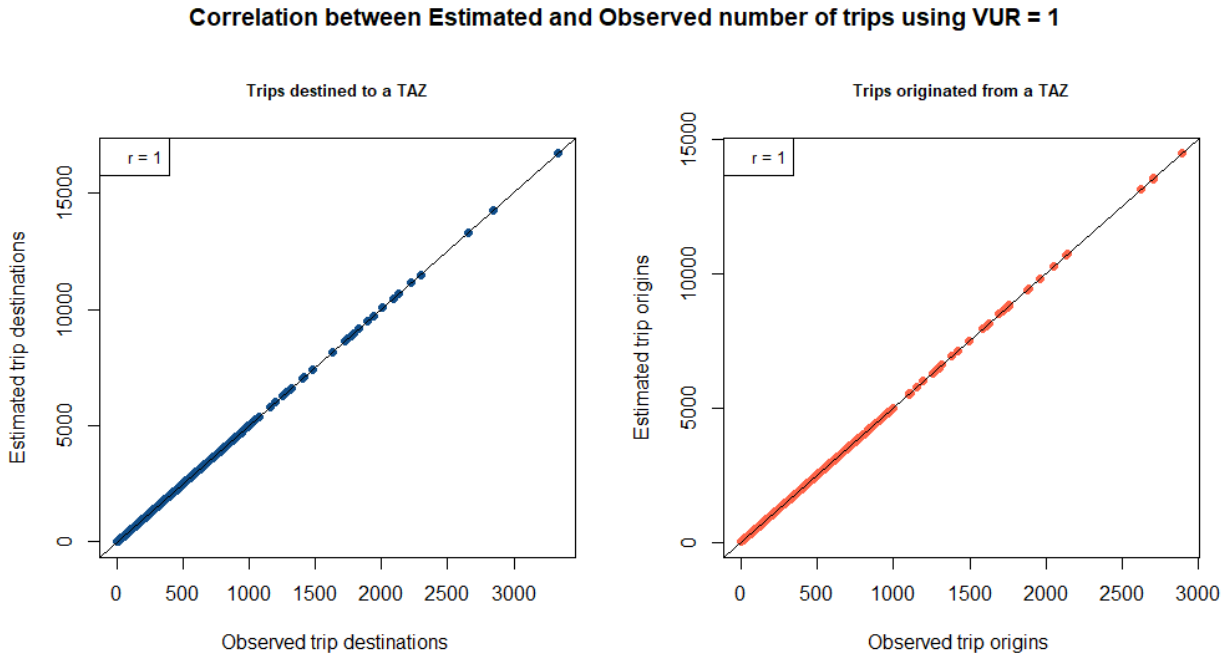


Figure 88. Graph. Correlation of estimated OD with observed trips using VUR = 1.

On the other hand, if one randomly varies VURs within different ranges for different TAZs, the perfect correlation disappears. The VURs are assumed to follow a uniform distribution. Two VUR ranges were tested: 0.2-0.8 and 0.3-0.5. For a given TAZ, its VUR is randomly generated from the specific range (0.2-0.8 or 0.3-0.5). With the increase of the range of VURs that are randomly assigned, the variation of the estimated OD (i.e., after upscaling) also increases, although the correlation coefficient still remains high. The variation in estimated trips originated from TAZs is found to be higher than that of the estimated trips destined to TAZs. The reason for this is probably due to the assumption that the trip origins are home locations. This is hardly true for the study area due to its small size. The VUR is used to randomly sample a fraction of the unique VIDs from each TAZ as vehicle trips. Changing the VUR randomly for different TAZs results in random changes in the observed number of vehicle trips originated from each TAZ. As the estimated OD table is finally upscaled by the total daily demand to match the PSRC demand model, the effect of the variation in the vehicle trips destined to TAZs is minimized. This leads to higher correlation between the observed trips destined to TAZs and the estimated total trips destined to TAZs.

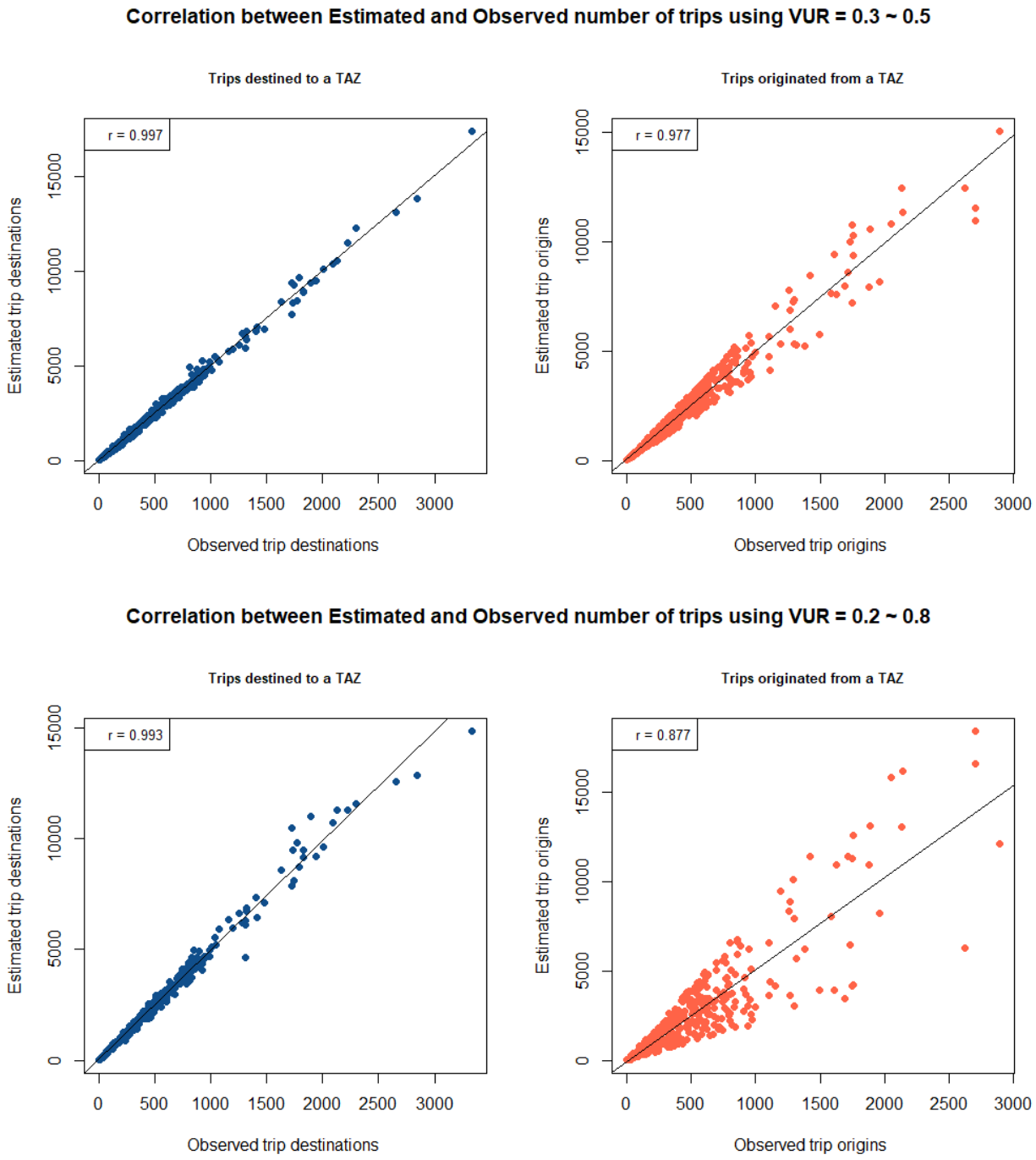


Figure 89. Graph. Correlation of estimated OD with observed trips using a range of VUR.

6.0 Comparisons and Observations and Implications to OD Analysis

In this section, the similarities and differences between the mobile phone and GPS data are discussed, as well as implications to OD analysis. Consistent with the structure in section 4.0 and section 5.0, the comparison starts with the underlying data generation process for the two datasets, then proceeds to discussing the zeroth, first, and second order properties. This section is concluded with a discussion on OD estimation.

6.1 Data generating process.

The mobile phone data, more specifically the sightings data, is generated from users' cell phone activities including for example, receiving/making a phone call, text messages and other cell phone network-required activities. Every sighting in the dataset is an instance at which a particular phone is seen on the network. It also means that when a sighting is generated, the user may be conducting an activity or a trip. For the vehicle-based GPS data, they are generated while the vehicles are in the trip-conducting process. Figure 90 illustrates these differences. In the figure, a hypothetical individual's one-day activity and travel pattern is illustrated during which the person leaves home at 7 AM, drives to a park and ride (P&R) station to ride the train to work. She arrives at work at 730 AM, has a work-based lunch tour around noon by walking, and then leaves work at 430 PM. Prior to going home, she arrives at a recreation place by train at 450 PM, spends 1 hour there and leaves there at 550 PM by train. She arrives at the P&R station at 610 PM, where she picks up her car and drives home, arriving at 640 PM. As shown, mobile phone signals can potentially show up throughout this one-day activity and travel pattern while the vehicle-based GPS data are only available when the individual drives her car. Figure 90 also illustrates the differences in spatial and temporal properties between the two datasets: spatially on the accuracy of location estimation, much larger amount of error are present for mobile phone data than for the GPS data; temporally, it is seen that though GPS signals only show up along the trip segment, they are more closely clustered together than mobile phone signals.

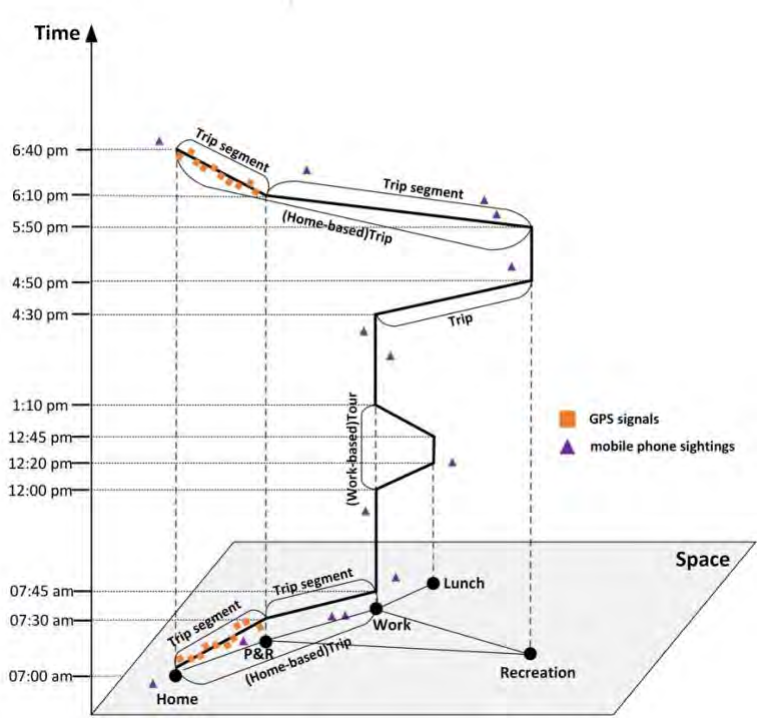


Figure 90. Graph. Illustration of mobile phone and GPS data on a hypothetical person's one-day activity and travel pattern.

6.2 Zeroth order properties

Locational accuracy. Location accuracy measures the distance in meters between the observed location and the actual (true) location. GPS data has much higher accuracy levels in location estimation than mobile phone data. As noted earlier, GPS data typically has an accuracy level of below 10 meters while the average for the mobile phone data is about 300 meters. Figure 12 in Section 4 shows that for the mobile phone data studied in this report, the mean is about 450 meters, meaning that a device can be found within 450 meters of the observed location with 90% probability. For this particular dataset, the error range is larger than the average reported for by the vendor, indicating the cell phone tower density for the Buffalo metropolitan area is likely lower than other areas for which the vendor owns similar data. The difference between the two datasets suggests that activity location or trip end capturing is likely much for accurate for GPS data than for the mobile phone data, though it is also important to note that high accuracy for location estimation is not necessarily required for OD analysis as they are typically done at the zone levels.

Weekly patterns. Both datasets show a clear and consistent weekly pattern: in general weekdays have more sightings than weekends and Sundays have the least. The consistency between the two datasets suggest that emerging datasets can play an important role in conducting trend analysis even on a real-time basis, an important capability that can be potentially realized quickly.

Temporal sparsity. Temporal sparsity refers to how sightings or GPS data records are distributed temporally. Three measures are developed to quantify temporal sparsity.

The first measure is simply the distribution of users with different number of days observed: in the mobile phone data, about 18% of the users are observed only one day and less than 1% of the users are observed at least once every day throughout the month-long study period (see Figure 5). For the GPS data, nearly 90% of the users are observed for only one day, followed by about 5% for two days. Only about 0.06% of the vehicles are observed every day for 30 days (see Figure 60).It is important to note that percentages for the two datasets are not directly comparable because the IDs in the mobile phone dataset remain the same for the entire 30-day study period while vehicle IDs in the GPS data are scrambled by the vendor periodically. This may create challenges if one wants to use the vehicle-based GPS data to infer certain activity locations (such as home locations).

The second measure we use is lifespan, defined as the difference between the last and first days an ID is observed. The similarity between the two datasets is that one day of lifespan has the largest portion of users: nearly 20% for the mobile phone data (see Figure 6) and nearly 90% for the GPS data (see Figure 53). Again, the actual percentage for the GPS data is likely much lower given that vehicle IDs are frequently scrambled.

The third measure is the time interval (in seconds) between consecutive sightings. Figure 1 and Figure 58 show the distribution of median intervals for the two datasets in log scale. Two observations can be drawn: 1) in both cases, the distributions for weekdays and weekends are very similar; 2) the distribution for GPS data follows a linear curve while that for mobile phone data is truncated when the interval exceeds 200 seconds. This reveals a key difference between the two datasets—the mobile phone data comprise many sightings that are clustered in time together as well as those that are far away from each other temporally, while the GPS data comprise primarily records that are clustered together.

Two implications may be drawn from analyzing the temporal sparsity of the two datasets. First, both datasets are likely to miss trips, but for very different reasons. Trip rates estimated from both datasets are likely under-estimated, as shown in Figure 33, Figure 34, Figure 72, and Figure 73. Moreover, the kinds of trips missed may be systematic: non-vehicle based trips and trips of longer duration for GPS data; morning trips and short trips for the mobile phone data. Second, the fact that many trips show up for only one day suggests that there is likely a large portion of users in the dataset who are visitors and passerby and they will need to be filtered out when comparing against household travel surveys that are resident based.

Time of day pattern. The two datasets exhibit different time of day patterns. The mobile phone data peaks only in the early afternoon between 1 PM and 3 PM and the patterns for weekdays and weekends are similar (see Figure 7); the GPS data however has two clear peaks on weekdays: in the morning around 9 AM and in the afternoon between 3 PM and 6 PM (see Figure 59). The weekend distribution has only one long afternoon peak from 1 PM to 6 PM. These observations suggest that mobile phone data cannot reliably capture traffic patterns. GPS data does a better job though the actual time periods for morning and afternoon peaks require further investigation.

6.3 First order properties

Locational uncertainty. Location uncertainty measures the amount of uncertainty associated with a location estimate. There is uncertainty in the location estimate in mobile phone data by nature—since every sighting is generated through triangulation of multiple cell towers, the location estimates for the same location at different times are different. Thus, the extent of locational uncertainty can be quantified by the amount of separation (or distance) between different sightings representing the same location. As shown in Figure 14, for mobile phone data, about 80% of the sightings belong to clusters that are less than 500 meters and the rest belong to larger clusters that are more than 500 meters. The location uncertainty of GPS data on the other hand, is much lower, as expected, given its high accuracy level in location estimation. These observations send a strong message to researchers using mobile phone data and it is: the data must be processed first before trips can be inferred. The recent study by Wang and Chen (2017) shows that without accounting for oscillation, regularity of human mobility patterns is over-estimated.

Activity duration distribution. In both datasets, activity durations are inferred and compared against those in the corresponding travel surveys (the Buffalo household travel survey for the mobile phone data and PSRC household travel survey for the GPS data). While the comparisons are largely similar in both cases, there are a few notable differences. The GPS data identifies about 85% of the durations of less than an hour corresponding to only about 45% from the PSRC travel survey (see Figure 67). The reverse order seems to be true for the mobile phone data—about 40% identified from the mobile phone data last less than an hour, corresponding to about 50% from the Buffalo travel survey (see Figure 19). On the other hand, for activities of longer duration (between 300 and 800 minutes), GPS data captures significantly less than those in the travel survey (see Figure 67). The reverse is true for the mobile phone data (see Figure 19). This difference suggests that the two datasets, due to their different data generating processes, tend to capture activities of different durations.

Spatial distributions of origins and destinations. For both datasets, the spatial distributions of trip origins and destinations are largely similar for both weekdays and weekends (see Figure 24 and Figure 25 for mobile phone data and Figure 68 and Figure 70 for GPS data). When compared against the results from the MPO model results and the census population data, the correlations with the MPO model results are much higher than with the census population and overall the mobile phone data has higher correlations than the GPS data (see Table 8, Figure 69, and Figure 71). Additionally, for mobile phone data, trip origins have higher correlations than destinations while the reverse is true for the GPS data. Again, the differences between the two datasets are caused by the nature of their respective data generating processes.

6.4 Second-Order Properties

Trip rate distribution. For both mobile phone data and GPS data, the estimated trip rates are lower than those obtained from survey data. In case of mobile phone data, the estimated mean trip rates (per day) are about 1.8 for weekends and 1.6 for weekends, while Buffalo travel survey data show

3.9 for weekdays (see Figure 33 and Figure 34). In the case of the GPS data, the estimated mean trip rates (per day) are 2.1 for weekdays and 1.7 for weekends, while the PSRC survey data show 4.4 for weekdays (see Figure 72 and Figure 73). The reasons underlying can be quite different. For mobile phone data, since the phone sighting data are not closely related to travel, thus missing those trips during which there are no phone usage. For the GPS data, the main reason is probably the small size of the area such that not all trips of a traveler (vehicle ID) can be captured for a given day in the dataset. Additionally, both datasets may contain non-residents passing through the region and they are likely to have low trip rates.

Departure/arrival times. Similar to the temporal patterns of the observations (0-order), the departure/arrival times show different patterns when comparing mobile phone and GPS data. Mobile phone data only capture the afternoon peak, completely missing the morning peak (see Figure 36 and Figure 37). The GPS data, especially the weekday data, can capture the morning peak reasonably well, but not so for the afternoon peak. Neither of them can capture the relatively low number of trips during the middle of the day (see Figure 76). Overall, the GPS data can capture the departure/arrival times slightly better by comparing with MPO survey results. This may be due to the fact that the GPS data were collected mainly from vehicles thus representing the vehicular travel patterns better. The size of the study area may also contribute to the discrepancy between arrival patterns obtained from GPS data and survey data. If more data had been collected from a much larger area, such discrepancy may be decreased.

Trip durations. Trip durations estimated from the two data sources show startling differences and details on how trips are derived are provided in Section 7. For the mobile phone data, about 50% trips estimated from mobile phone data have a trip duration larger than 60 minutes, while Buffalo travel survey data show only 1% of trips lasting longer than 60 minutes (see Figure 39). For GPS data, however, the estimated trip durations concentrate on those less than 15 minutes (nearly 88%), while survey data show about 60% for such trips (see Figure 78). The difference may again be attributed to the inherent differences of the two datasets: mobile phone data are generated through phone use activities, while GPS data are generated from in-vehicle devices or mobile apps but for a much smaller area. The underestimated trip rate from mobile phone data is also consistent with the larger trip duration: multiple actual trips might have been considered as a single trip when processing the mobile phone data, thus leading to much longer trip durations. For GPS data, the smaller trip rates are mainly due to the smaller area, which also tends to capture trips with shorter durations more than trips with longer durations.

Zero cells. Regarding zero cells, the trends estimated from the two data sources are similar (see Figure 44, Figure 45, Figure 80, and Figure 81). More zero cells are located on the boundary zones of the study area, probably due to the fact that boundary zones tend to generate/attract more trips from zones outside the study area, which cannot be captured by the datasets. Furthermore, as data from longer time periods (i.e., more days) are available, the percentage of zero cells decrease monotonically. However, even with the longest time periods (30 days for mobile phone data and 90 days for GPS data), the estimated OD trips still show considerable zero cells: about 30% for mobile phone data (see Figure 47) and 65% for GPS data (see Figure 79).

OD collinearity. OD collinearity indicates the correlation between the observed trips (observed because they are directly inferred from the mobile phone or GPS data) and the upscaled OD

demands (also from the same datasets). This measure shows primarily the features of the OD upscaling methods that expand the observed trips to the regional totals obtained from MPOs model results (this is to account for those trips that are missed from the mobile phone or GPS data). It is assumed that very high correlations show almost linear relations between the observed trips and the estimated ODs, and are thus less desirable (or at least the upscaling methods are probably too simplistic). Unfortunately, the correlations computed from both mobile phone data and GPS data are very high, on the order of 0.9-1.0, representing an almost linear relation between the observed trips and the estimated OD demands (see Figure 48 and Figure 88). When changing some of the upscaling parameters, such high correlations may get reduced significantly (see Figure 89). However, this will require obtaining zone-specific parameters such as zone-level penetration rates of GPS devices or identifying more precise filtering rules to apply for mobile phone data. Neither of these however can be easily decided in real world applications. The results thus suggest that there is much work to be done before one can use emerging datasets such as mobile phone and GPS data for direct OD analysis.

OD matrices compared with MPO OD demands. By comparing the estimated OD demand matrices from the two data sources and the MPO demand matrices, the results show correlations in the range of 0.6-0.7 (see Figure 49 and Figure 87). This indicates that overall, the estimated OD demands do not represent the MPO demand matrices well, even though they may be able to produce OD demand matrices. Notice here that MPO demand matrices are also estimated from survey and other types of transportation data, and are thus not the "ground truth." We assume that they are more representative of the true OD demand matrices since the data collection and estimation processes are more controlled and carefully designed. However, given the mediocre correlations between big-data-estimated OD matrices and MPO demand matrices, more research and investigations are needed to further study the data properties and develop more sophisticated OD estimation methods to produce better representative OD matrices from the big data sources.

6.5 Implications to OD Analysis

In summary, specific characteristics of transportation big data may contribute to the different properties of the analysis results, as discussed above, when using the data for OD related analysis. These characteristics may be categorized as follows:

1. *Updating frequency of the random vehicle or device ID*. Raw data are rarely available and each processed data record is often associated with a random ID. How often this random ID is updated has important implications to OD inference. If the ID is updated frequently, say once every few hours (which is the case for the GPS data but not for the mobile phone data used in this research), trips longer than that duration cannot be captured, leading to underestimation of trip rates, trip durations, and OD demands. Even if the IDs are updated every day or every few days, algorithms will have difficulty of correctly identifying home and work locations.
2. *Location accuracy and uncertainty.* Low location accuracy can lead to high location uncertainty, which will pose challenges when identifying trip ends and consequently affect OD estimation. On the other hand, high location accuracy, although helpful for trip end identification and related analysis, may also introduce challenges due to other possible considerations such as data privacy/security. This is not to say that mobile phone data,

due to their high locational uncertainty, are free of privacy issues. In fact, the combination of longitudinal data nature and high level of regularity exhibited in human mobility patterns, suggests that reasonable accurate location estimate can be made from the mobile phone data (Chen et al 2014; Gonzalez et al 2008).

3. *Size of the study area.* The above comparisons clearly indicate that the size of the study area is crucial. If the area is too small (like the GPS dataset used in this research), the data cannot capture the complete trip of a traveler (or vehicle ID), leading to significant underestimation of some of the properties such as trip rates and trip durations. Therefore, for OD-related analysis, data from a relatively large area (such as an entire city or a region) should be used.
4. *Duration of the collected data.* Both datasets show clear weekly patterns and daily variations. As a result, for OD-related analysis, data from at least a few weeks (e.g., 3 weeks of data are needed for mobile phone data) should be analyzed (see Figure 30 and Figure 31) though the representativeness of the OD demands may still be a problem. For trend analysis, data of longer durations are required.
5. *The data generation process.* Although in general it is hard, if not impossible, to know exactly how the data are generated, even some basic investigation and understanding of the data generation process may be helpful. For example, mobile phone data are generated mainly for phone activities, which may or may not be trip related. GPS data, if collected from vehicle navigation devices or monitoring systems, are more likely related to vehicular travels. Such knowledge is relatively easy to obtain, which can nevertheless provide useful insight and explanations to the patterns of the properties discovered from the data.

It is expected that the various characteristics analyzed in this study are also relevant to other emerging datasets (e.g., app based data) that can be potentially used for planning purposes such as OD estimation. Issues that arise from these characteristics all converge to the fundamental question: does the dataset at hand represent the travel patterns of a region? The issue of *representativeness* should be the central question when analyzing and using big data for OD related analysis. The very first task in answering this question is to obtain a thorough understanding of the data at hand. As shown in this study, the two datasets have very different characteristics due to their respective data generation processes. The framework (zeroth, first, and second order properties) proposed in the study can be readily applied to other emerging datasets for this purpose. In the long term, future research is critically needed to develop tractable methods to address those issues to produce more accurate and reliable trip/OD related information.

7.0 Appendices

7.1 Appendix A. Processing Mobile Phone Data

In this appendix, methods used to address two problems associated with mobile phone data (locational uncertainty and oscillation) are provided. These methods follow the data-processing framework proposed by Wang and Chen (2017). As is put in the Section 4.2.2, locational uncertainty refers to the issue that every location estimate in the data (even estimates for a single location) is unique, as the results of triangulation. And the oscillation issue is pure signaling related: some location changes are recorded due to signaling activities, as opposed to users' movements.

7.1.1 A.1. Addressing Locational Uncertainty

Due to the locational uncertainty, location records for a potential activity location appear distinct in their expressed pairs of latitude and longitude. Putting these location records together during the observation period, one would find them closely distributed in space. Based on this observation, an incremental clustering algorithm is first applied to reveal activity locations either visited only once or multiple times (Figure 91). For each user, traces in the entire observation period are put together and clustered without regarding their time ordering. This technique identifies common activity locations by aggregating traces that are close in space but may be far away in time (e.g., several days). Following the clustering, **an activity location is represented by the centroid of a cluster when duration exceeds a given threshold T_c (set as five minutes).** The duration of each visit is the largest time difference of consecutive traces that belong to one cluster.

```
Input: multiple-days traces of one user d_list;
Output: A set of clusters of traces;
Steps:
    gather traces of multiple days to d_list;
    C_set ← {};
    create the first new cluster C_new and add any one trace d_0 to C_new;
    add C_new to C_set;
    C_current ← C_new;
    for other traces d_i in d_list not belonging to any cluster:
        if distance(d_i, C_current) < R_c:
            add d_i to C_current;
        else:
            C_current ← None;
            for all cluster C in C_set:
                if distance(d_i, C) < R_c:
                    add d_i to C;
                    C_current ← C;
                    break;
            if C_current =None:
                create new cluster C_new;
                add C_new to C_set;
                C_current ← C_new;
    output C_set;
```

Figure 91. Illustration. Incremental clustering algorithm.

The above clustering method requires a spatial constraint Rc as an input, which is found via trial and error. Since the locational uncertainty directly hinders the identification of activity locations,

different values of Rc differently affect the number of activity locations identified. Figure 92 gives the average number of distinct activity locations visited per day per person n_{AL} as a function of Rc. It suggests the suitable Rc appears at 3280 ft (1 kilometer). When Rc is small, one actual activity location may be split into several clusters. With increasing Rc, small clusters start to merge into meaningful representations of actual activity locations. Above 1 kilometer, it is possible that some clusters may absorb passing-by traces (those generated during a trip) and lengthen their duration, thereby meeting the temporal constraint Tc, consequently leading to the growth of n_{AL}. This is witnessed by the slight increase of n_{AL} and by the steep rise of mean cluster duration if Rc exceeds 1 kilometer (Figure 92).

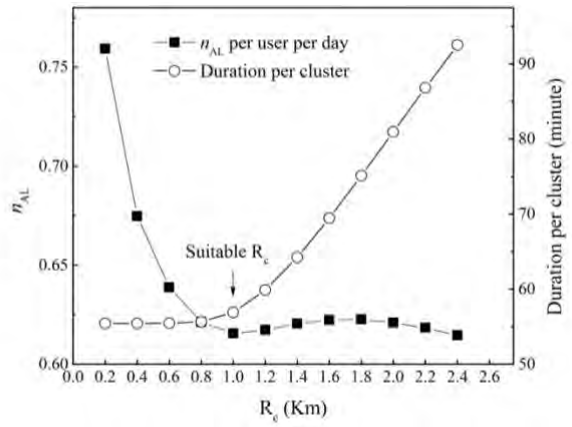
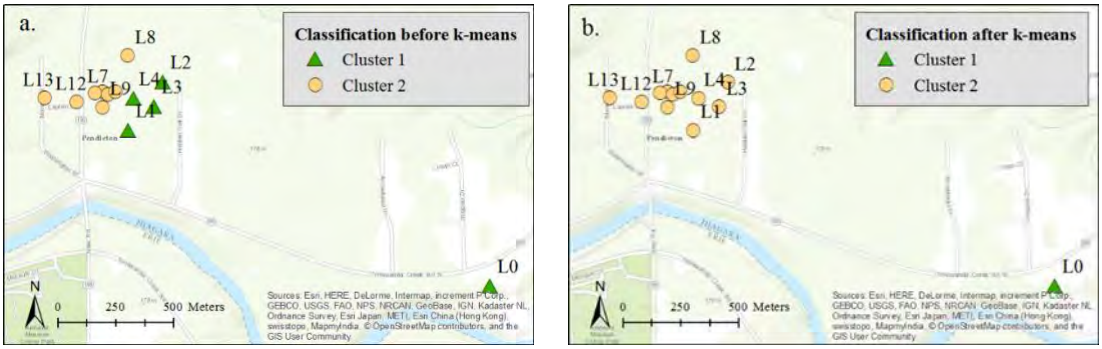


Figure 92. Graph. Number of distinct activity locations and mean cluster duration as a function of Rc.

An order problem is not handled by the incremental clustering algorithm: clustering results are subject to the order of how traces are clustered, which can result in unreasonable clusters. Figure 93-A gives one example. The incremental clustering algorithm yields two strange clusters if a new cluster is created at L0. This order problem is tackled with the k-means clustering algorithm. More specifically, one initializes the k-means clustering algorithm using the number of clusters and centroids of clusters that are yielded from the incremental clustering algorithm. The k-means algorithm resorts a trace to the cluster with the nearest distance to the centroid of the cluster. The clustering error in the example is corrected (Figure 93-B).



A. The problematic clustering. B. Corrected clustering.

Figure 93. Illustration. Illustration of an order problem.
Source: OpenStreetMap

7.1.2 A.2. Addressing the Oscillation Problem

A time-window-based method is applied to detect traces generated due to the occurrence of the oscillation phenomenon. It scans trajectories with a short time window T_w, which always starts at the last trace of a cluster and returns a sequence of traces. Among the sequences returned, the one containing at least one circular event is considered as an oscillation sequence. A circular event refers to a tour when one device starts its connection with one location L_0, later jumps to distinct locations, and returns to location L_0. Since it is less likely for any user to make a tour within a short time window, the oscillation sequences detected may contain traces from oscillation phenomenon.

A suitable T_w is determined via trial and error. Figure 94 shows the oscillation ratio as a function of T_w. The oscillation ratio is defined as the ratio of the number of detected oscillation traces over the total traces. The elbow rule[7] suggests T_w = 5 minutes as a reasonable choice to separate oscillation cases from real trips.

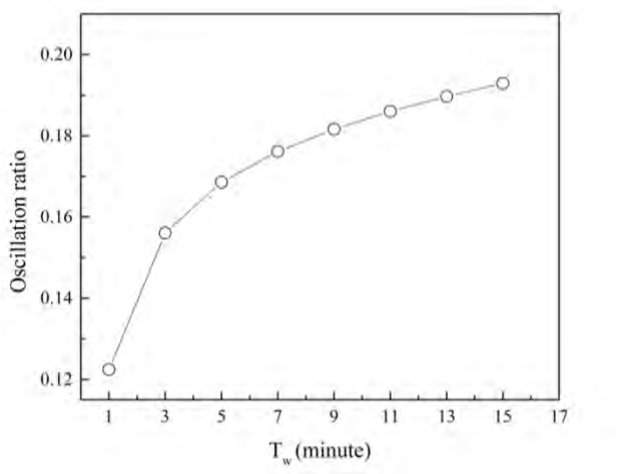


Figure 94. Graph. Average oscillation ratio as a function of Tw.

In some cases, the time window may be too short to capture the circular event even though the oscillation occurs. To detect these cases, a modified pattern-based method is applied. The pattern-based method is modified as it alone may mistake actual trips for oscillation. For each sequence detected with a twice switches pattern (i.e. the device jumps to one faraway location then returns), one looks for additional evidence: only if at least one time interval, between any two consecutive traces in the sequence, is found shorter than T_w, it is identified as an oscillation sequence. For each oscillation sequence detected, oscillation traces are removed and replaced with a meaningful location, which is identified as the one where the user spends most of his/her time during the entire observation period.

7.1.3 A.3 Literature Review of Route Detection

When longitudes and latitudes of mobile phone locations are not available, the chronically ordered sequence of cell tower IDs can be matched with the sequences of the cell towers to derive

[7] The elbow rule is a method to help finding the appropriate number of clusters in a dataset.

possible travel routes (Laasonen, 2005). However, if longitudes and latitudes are available, the travel routes of a mobile phone users can be mostly detected using map-matching methods. In doing so, the activity locations (trip ends) are first estimated, and their nearest transportation nodes are selected as the origin and the destination if a user visited the two activity locations (also called the stationary locations) consecutively. Then, this pair of origin and destination location is superimposed on a transportation network map, and the route is roughly estimated (Bayir, Demirbas, & Eagle, 2010; Chen et al. 2016; Leontiadis et al. 2014; Tettamanti, Demeter, & Varga, 2012). The accuracy of this process can be improved if there are some intermediate locations (also called passing-by locations or mobile locations) because the chronically ordered sequence of these locations, and their closeness to physical road networks also help detect travel routes (Iqbal et al. 2014a; Laasonen, 2005). However, in areas with dense road networks, route detection can be relatively inaccurate (Chen et al. 2016; Tettamanti et al. 2012; Tettamanti & Varga, 2014).

Since GPS data contain longitude and latitude information, route information can be more easily detected using GPS data. This typically involves positioning GPS locations onto a map using certain map-matching techniques. Route information can then be obtained accordingly, as discussed above. More details are omitted here.

7.2 Appendix B. Processing GPS Data

7.2.1 B.1. Trip Information Extraction from GPS Data

If analyzing the GPS observations of a vehicle in a time series on a map, one may find those GPS observations often show two geographic patterns. One is that the time-consecutive GPS points distribute along a certain route with nearly equal distances between each other, which indicates a smooth movement of the vehicle, while the other one is that several GPS points aggregate within a small area, showing that the vehicle stops for a while or wanders within this area. In most cases, the GPS observations of one vehicle often alternately present these two types of geographic distributions, thereby proving that the activities of a vehicle in one day usually contain both movements and stays (stops) due to the trips the vehicle has made. The starting and ending locations of a trip correspond to the stays. However, not every stay corresponds to a trip end. Some stays may be due to non-trip-related stops such as traffic stops at a red signal. One of the key challenges of trip information extraction from GPS data is to distinguish trip ends from other types of stays.

The trip information extraction process mainly consists of three steps. The first step uses a trip end identification algorithm to differentiate trip ends with other sorts of stays, and produces a trip-related stay table. Next, the method executes the trip reconstruction process to extract raw trip table based on the stay table obtained in the previous step. In the third step, the speed condition and activity reconstruction process are incorporated to generate estimated trip table and activity information, respectively. The flow chart and corresponding algorithms of the trip information extraction method are given in Figure 95.

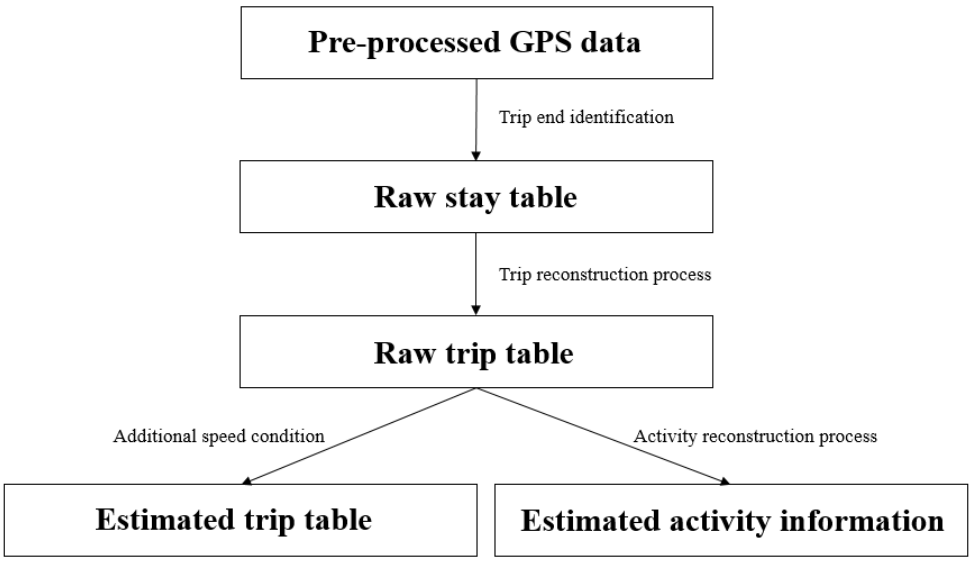


Figure 95. Chart. Flow chart of trip information extraction method.

7.2.2 B.2. Trip End Identification Using GPS Data

In this project, a spatial-temporal clustering approach (Ye et al. 2009) is utilized to process GPS data to reconstruct trip information, and then to identify trip ends. The algorithm identifies trip ends by inferring stay points. A stay point is a geographic area where a vehicle stays for a certain length of time. Figure 96 shows how this algorithm works for stay point inference. Assume a GPS trace was captured for a vehicle. The leftmost point is the beginning of a series of GPS points of the vehicle. After that, the vehicle was captured four times within a spatial region (within a distance threshold) for a period of time (larger than a predefined time threshold). Then the algorithm treats this cluster of GPS points as a stay (or stop). The average latitude and longitude of those points act as the coordinates of the inferred stay point, and the times of the first and last GPS point of this cluster are the beginning time and ending time of this stay, respectively. The identified stay points will be considered as the trip ends, and trips can be constructed accordingly.

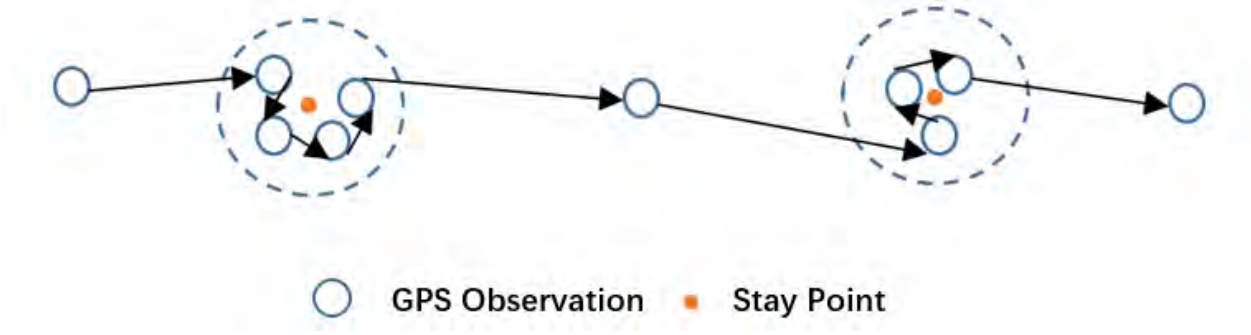


Figure 96. Illustration. Illustration of trip end identification.

7.2.3 B.3. Trip Reconstruction Process

Once stay point information is obtained, GPS observations and the stay tables can be combined to produce trip information. The idea to generate the trip table stems from the fact that two stays are connected by one trip, and one trip end (stay point) is usually the joint of two adjacent trips. By connecting two consecutive stay points and regarding the movement of vehicles between these two points as a trip, one may generate the trip tables. However, this method only works for those trips produced when traveling between two stay points. For movements that begin/end with a single GPS point (not an identified stay point), this trip reconstruction idea might leave out certain type of trips. In the trip reconstruction process, a new step is added to avoid this situation. If the first GPS observation is more than 656 ft (200 meters) away from the first stay point, a new trip which begins with the first GPS observation and ends with the first point of the stay would be identified as a new trip. The same rule applies to the situation involving the last stay and the last GPS observation. Figure 97 illustrates the process of reconstructing the raw trip information.

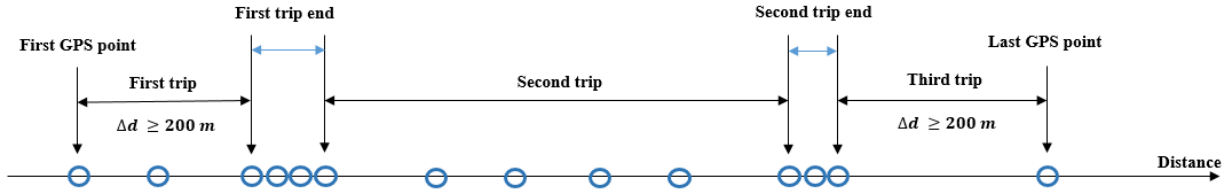


Figure 97. Illustration. Illustration of trip reconstruction process.

7.2.4 B.4. Extraction of Estimated Trip and Activity Information

The trip and stay points identified by the first two steps suffer obvious flaws when applied to GPS data. First, the trip information from Step 2 might not be complete and highly unreliable since the speed condition is not considered. It is well accepted that a complete trip by vehicles often begins and ends with a zero (or low) speed. However, the approach discussed above does not take speed into consideration. Since the research area in this project is small, there are many trips originating/destining outside the area but ending/starting inside it. Under those cases, the GPS dataset collected within this area cannot capture all information of an entire trip. For example, if a vehicle originating outside of the area travels through Interstate 5 (a freeway traversing the Seattle metropolitan area) and finally arrives at downtown Seattle, the dataset just covers the observations from the area boundary to the destination. Even if those observations could contribute a stay point and a trip, this trip would not be a true OD trip, since the data shows it originates from the boundary zone instead of its true origin. In order to solve this problem, available speed information provided by the GPS dataset is considered. A raw trip, only if the speed of its two trip ends is less than 10 Mile Per Hour (MPH), would be treated as a (estimated) true trip. If the trip end is a single GPS observation, its speed is the speed of this GPS observation. If the trip end (stay point) consists of many GPS points, its speed equals the mean value of those points.

Second, this algorithm identifies trip ends only based on the spatial-temporal relationship between GPS observations in the first step and does not include the relationship between trips and activities. As a result, the derived stay duration would not be the true activity duration if a vehicle only produces one trip. For instance, if a vehicle is observed to produce only a trip in one day and

this trip begins and ends with a GPS point cluster, those two identified stay points cannot be counted as true activity locations because they just indicate starting/ending stage of a trip rather than true activities. In other words, some essential attributes, such as activity beginning time, ending time, duration, are missing. In this project, if a vehicle stays at a location for some time, it is assumed that the corresponding driver or passengers are doing some activities there. To obtain more reliable activity information, an additional step is proposed to process the raw trip table. This step relies on the basic assumption that a true activity is always connecting two trips. If there is only one trip for a vehicle, then one cannot infer activity information. Under such assumption, if the distance between the ending point of a previous trip and starting point of a later trip is close enough (i.e., a distance threshold of 656 ft (200 meters), these two points will be treated as one activity location and the activity duration can be computed accordingly. In sum, the pseudocode of the proposed trip information extraction method can be shown in Figure 98.

7.2.5 B.5. Discussion about the Selection of Thresholds

In the trip end identification algorithm, only two parameters need to be determined before processing the GPS observations: the time threshold and the distance threshold. As discussed in the literature review, duration (i.e., the dwell time) and other threshold-based methods are popular approaches of detecting stay point locations using GPS data. For example, different values have been considered as the threshold of dwell time, ranging from 45 to 900 seconds, such as 45, 120, 180, 200, or even 900 seconds. However, few studies further tested various distance thresholds for trip-end identification. These values depend on the local transportation situation and characteristics of the GPS data. If the time threshold value is too small, a temporary stop (i.e., stops caused by traffic congestion) may be mistakenly identified as an activity location. If the distance threshold is too small, one actual stay may be split into two separate stays. On the other hand, if the time threshold or distance threshold is too large, multiple stays may be treated as one stay. In this project, 328 feet and 300 seconds are selected as the distance threshold and time threshold.

In this project, different values of time threshold are tested to see how they affect the final output (the number of stays and the number of trips) by using the proposed algorithm. One-day (01-15-2017) GPS data was chosen as an example and the results are shown in Figure 99 and Figure 100. As the time threshold value increases, both the number of stays and the number of trips decrease. It is obvious that a longer time threshold may combine two temporally adjacent stay points into one stay point, thereby directly reducing the numbers of stays and trips. In some cases, excessive large threshold values may even result in critical information loss. By looking at the curve of the number of stays vs. time thresholds, one can easily find that when the time threshold lies in the interval of [120, 300], the decreasing rate of the curve is the largest, which reaches the minimum value near 300 seconds. After 300 seconds, the decrease rate becomes large again. This trend indicates that 300 seconds may be an appropriate time threshold for identifying stays.

Trip End Identification Algorithm

Input: A set of GPS observations of one vehicle P

Output: A set of stay points SP

$i = 1$; pointNum $=$ dim(P); //the number of GPS observations
$timethreshold = 300; disthreshold = 100;$
while $i < pointNum$ **do**
  $j = i + 1;$
  while $j < pointNum$ **do**
    $dd = GPSdistance(p_i, p_j)$; //calculate the distance between two points
    if $dd > disthreshold$ **then**
      tt $= p_i.T - p_j.T$; //calculate the time difference between two points
      if $tt > timethreshold$ **then**
        $S_{coord} = Mean(p(k)_{coord} | i \le k \le j)$
        $S_{arrT} = p_i.T; S_{leavT} = p_j.T; S_{speed} = Mean(p(k)_{speed} | i \le k \le j);$
        $SP.insert(S);$
        $i = j;$ break;
      $j = j + 1;$
return SP.

Trip Reconstruction Algorithm

Input: A set of GPS observations of one vehicle P, a set of stay points SP

Output: A set of raw trip data TS

$NumofStay = dim(SP)$; //the number of stay points
if $NumofStay >= 1$ **then**
  $InitialDist = GPSdistance(p_1, sp_1); EndDist = GPSdistance(p_{pointNum}, sp_{NumofStay});$
  if $InitialDist > 200$ **then**
    $triplist = combine((p_1, sp_1); TS.insert(triplist));$
  if $NumofStay >= 2$ **then**
    for nn in 1: $(NumofStay - 1)$ **do**
      $triplist = combine((sp_{nn}, sp_{nn+1}); TS.insert(triplist));$
  if $EndDist > 200$ **then**
    $triplist = combine((p_{pointNum}, sp_{NumofStay}); TS.insert(triplist));$
else do
  $triplist = combine((p_1, p_{pointNum}); TS.insert(triplist));$
return TS.

Method for True Trip Information and True Activity Information

Input: A set of raw trip data TS

Output: A set of true trip data TTS, a set of true activity data TAS

$NumofTrip = dim(TS)$; // the number of raw trips
$speedthreshold = 10$; // the speed threshold for identifying true trips
for i in 1: $NumofTrip$ **do**
  $O_{speed} = ts_{i(O_{speed})}; D_{speed} = ts_{i(D_{speed})}$
  if $(O_{speed} \le speedthreshold$ && $D_{speed} \le speedthreshold$) **then**
    $TTS.insert(ts_i)$
for j in 1: $(NumofTrip - 1)$ **do**
  $dis = GPSdistance(ts_{j(D_{coord})}, ts_{j+1(O_{coord})});$
  if $dis \le 100$ **then**
    $activity_{coord} = Mean(ts_{j(D_{coord})}, ts_{j+1(O_{coord})}); activity_{duration} = ts_{j+1(O_{time})} - ts_{j(D_{time})};$
    $TAS.insert(activity)$
return TTS, TAS

Figure 98. Illustration. Trip information extraction method.

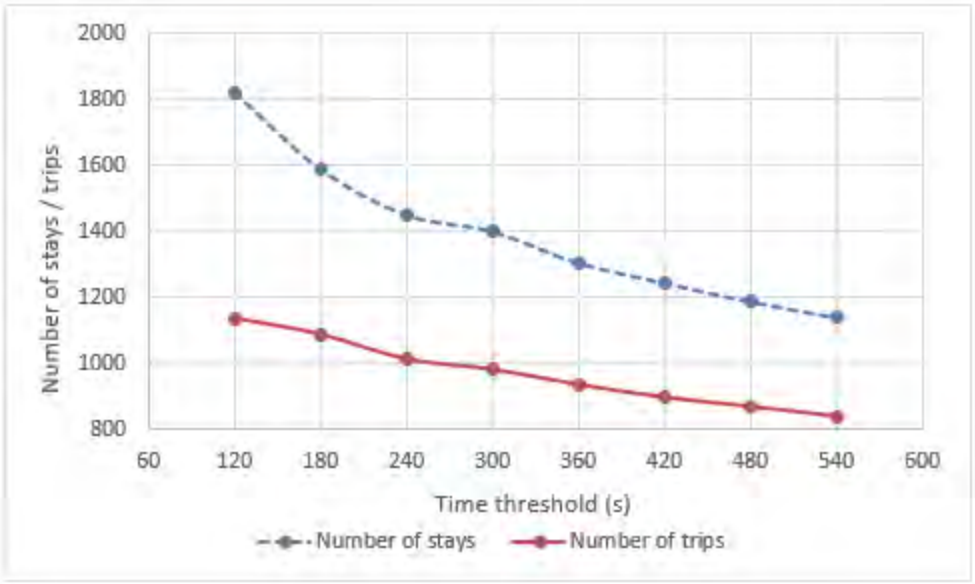


Figure 99. Graph. Number of stays/trips as a function of time threshold.

In addition to the time threshold, the distance threshold is investigated, which also plays a significant role in the spatial-temporal clustering method. Similar relationships between numbers of stays and trips vs. the distance threshold are presented in Figure 100. It can be found that in the intervals of [164, 328] ft ([50, 100] meters) and (492, 556) ft ([150, 200] meters), the decreasing rates of the curve are relatively small. Therefore, these two intervals might be good candidates for the distance threshold. The 328-ft (100-meter) distance threshold is selected in this project.

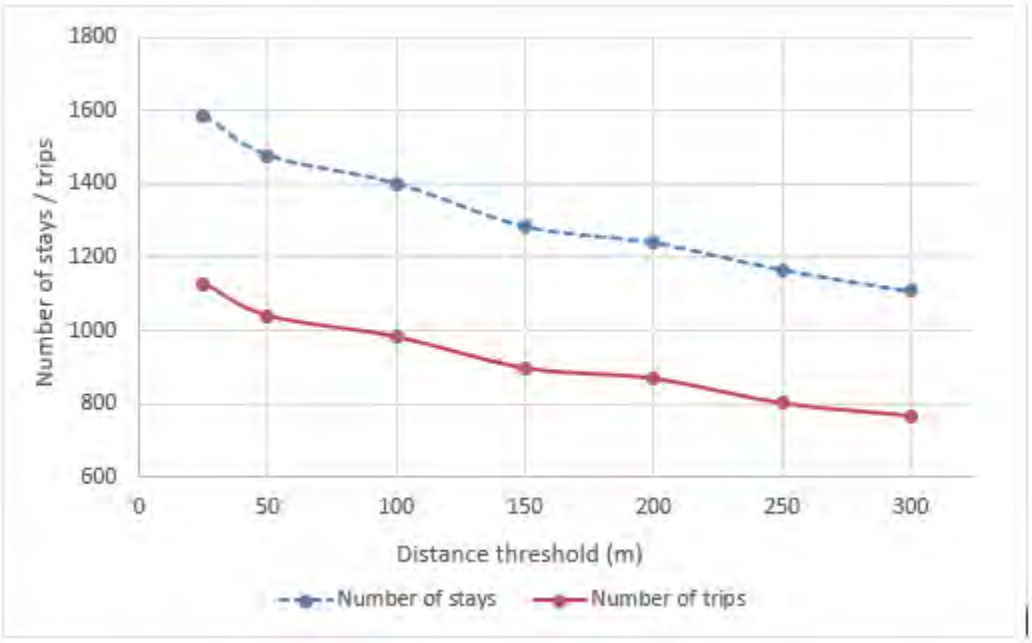


Figure 100. Graph. Number of stays/trips as a function of distance threshold.

Furthermore, the trip and stay results produced by two distance thresholds (328 ft (100 m) and 492 ft (150 meters)) are compared to further investigate the difference of the two distance thresholds. Figure 101 is an example. GPS observations are marked as blue points in the map in a temporal order (1 to 15). If the 328-ft (100-meter) threshold is used, it can generate two trips and one activity location (marked using red text). However, if the 492-ft (150-meter) threshold is used, it only keeps Trip 1, while Trip 2 and the activity are discarded. This happens because the large distance threshold could treat two or more stay points as one if they are close enough. In this case, the 492-ft (150-meter) threshold treats GPS point 12 as part of the stay and constitutes a larger cluster (the yellow circle). Once applying the trip construction process, the distance between Point 15 and this cluster is less than 656 ft (200 meters), which cannot satisfy the rules to construct a new trip, hence discarding Trip 2. The slight change of distance threshold even leads to the loss of such crucial information (i.e., one trip and one activity in this example). Therefore, the choice of the threshold requires careful consideration, possibly by studying the features of GPS data, geographical characteristics, and local transportation conditions.

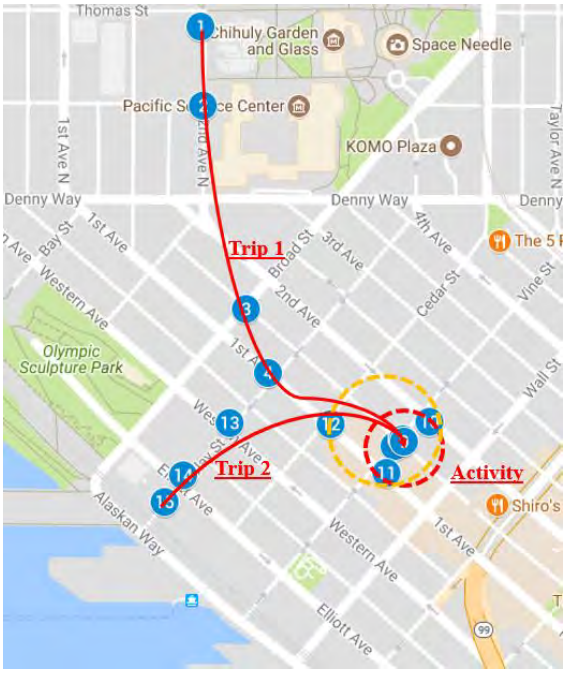


Figure 101. Illustration. Example of trip identification results by various distance thresholds.

Source: © Google

7.3 Appendix C. Pseudocode for OD Estimation from GPS Data

Finding the transient OD $(t - OD)$ according to Wang et al. (2012)

Wang et al. (2012) defined the transient origin-destination (t-OD) as the observed initial and final location of the user. In the GPS data, a user's location will be unavailable when he/she does not use his/her phone. Thus, there is a possibility to lose a segment of the trip information. Even if

only the transient origin and destination are captured with the phones, this still captures a large portion of the road usage.

This study made some modifications to the original algorithm to fit the need of GPS data. These modifications permit use of this method on irregularly sampled GPS data and comparison of the estimated results with MPO's OD table. The steps of the OD upscaling process are briefly described below.

1. Calibrate the total number of trips between OD pair (i, j), F_{ij} with the population in a zone:

$$F_{ij}^{all} = \sum_{n=1}^{N_k} T_{ij}(n) \times M(k)$$

Figure 102. Equation. Definition of total number of trips between OD pair.

Where, N_k is the number of users in zone k and $T_{ij}(n)$ is the total number of trips that user n made between zone i and zone j in the observational period.
and,

$$M(i) = \frac{N_{pop}(i)}{N_{user}(i)}$$

Figure 103. Equation. Definition of *M*.

Where, $N_{pop}(i)$ = Population in zone i.
$N_{user}(i)$ = Number of selected mobile phone users in zone i.

2. Calculate trips generated by vehicle users:
   1. Calculate vehicle usage ratio (VUR) in a zone

People use different transportation modes throughout their trips. Possible transportation modes include car (drive alone), carpool, public transportation, bicycle and walk. Calculate the vehicle using rate (VUR) in a zone as follows:

$$VUR = P_{car\ drive\ alone} + \frac{P_{carpool}}{S}$$

Figure 104. Equation. Definition of *VUR*.

Here, $P_{car\ drive\ alone}$ and $P_{carpool}$ are the probabilities that residents drive alone or share a car. According to Seattle Commuter Survey, 2014, 31.1% of workers drive alone to work and 9% used carpool (2). These values were used as the respective probabilities. The average carpool size was assumed to be 2.5. So,

$$VUR = 0.31 + \frac{0.09}{2.5}$$
$$= 0.346$$

2. Using the VUR calculated for each zone, randomly assign the transportation mode (vehicle or non-vehicle) to the users living in each zone.
3. Calculate the total number of trips generated by vehicles.

$$F_{ij}^{vehicle} = \sum_{v=1}^{V_k} T'_{ij}(v) \times M(k)$$

Figure 105. Equation. Definition of total number of trips generated by vehicles.

Where user v is a vehicle user, V_k is the number of users in zone k.

3. Calculate $t - OD$:

$$t - OD = W \times \frac{F_{ij}^{vehicle}}{\sum_{n=1}^{A} F_{ij}^{vehicle}}$$

Figure 106. Equation. Calculation of transient OD.

Where, W = Daily trip production for the entire population
A = Number of zones

8.0 References

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

Anderson, I., & Muller, H. (2006). Context awareness via gsm signal strength fluctuation. na. Retrieved from http://www.cs.bris.ac.uk/publications/Papers/2000528.pdf

Ban, X. J., Hao, P., & Sun, Z. (2011). Real time queue length estimation for signalized intersections using travel times from mobile sensors. Transportation Research Part C: Emerging Technologies, 19(6), 1133-1156.

Ban, X., Herring, R., Hao, P., & Bayen, A. (2009). Delay pattern estimation for signalized intersections using sampled travel times. Transportation Research Record: Journal of the Transportation Research Board, (2130), 109-119.

Ban, X., Sun, Z., Yang, X., Wojtowicz, J., Holguin-Veras, J., 2013. Urban Freight Performance Evaluation Using GPS Data. Chapter 5 of the Final Report on Off-Hour Deliveries submitted to Federal Highway Administrations (FHWA), US Department of Transportation (USDOT).

Ban, X., Wang, C., Kamga, C., Wang, X., Wojtowicz, J., Klepadlo, E., & Mouskos, K. (2014). Adaptive Traffic Signal Control System (ACS-Lite) for Wolf Road, Albany, New York. Transportation Research Board: Journal of the Transportation Research Board, No. C-10-13, Washington, DC: Transportation Research Board of the National Academies.

Bar-Gera, H. (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: Emerging Technologies, 15(6), 380–391.

Barth, M., Johnston, E., & Tadi, R. (1996). Using GPS technology to relate macroscopic and microscopic traffic parameters. Transportation Research Record: Journal of the Transportation Research Board, (1520), 89-96.

Bayir, M. A., Demirbas, M., & Eagle, N. (2010). Mobility profiler: A framework for discovering mobility profiles of cell phone users. Pervasive and Mobile Computing, 6(4), 435–454. https://doi.org/10.1016/j.pmcj.2010.01.003

Bernardin Jr, V. L., Ferdous, N., Sadrsadat, H., Trevino, S., & Chen, C.-C. (2017). Integration of National Long-Distance Passenger Travel Demand Model with Tennessee Statewide Model and Calibration to Big Data. Transportation Research Record: Journal of the Transportation Research Board, (2653), 75–81.

Bindra, S. (2016). Using Cellphone OD Data for Regional Travel Model Validation. Presented at the Transportation Research Board 95th Annual Meeting Transportation Research Board. Retrieved from https://trid.trb.org/view.aspx?id=1394422

Bohte, W., & Maat, K. (2009). Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. Transportation Research Part C: Emerging Technologies, 17, 285-297.

Bonnel, P., Hombourger, E., Olteanu-Raimond, A.-M., & Smoreda, Z. (2015). Passive Mobile Phone Dataset to Construct Origin-destination Matrix: Potentials and Limitations. Transportation Research Procedia, (11), 381–398.

CA.gov, (2002). 2002 Highway Congestion Data. Retrieved from http://www.dot.ca.gov/d4/highwayoperations/

Calabrese, F., Colonna, M., Lovisolo, P., Parata, D., & Ratti, C. (2011). Real-Time Urban Monitoring Using Cell Phones: A Case Study in Rome. IEEE Transactions on Intelligent Transportation Systems, 12(1), 141–151. https://doi.org/10.1109/TITS.2010.2074196

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

Calabrese, Francesco, Di Lorenzo, G., & Ratti, C. (2010). Human mobility prediction based on individual and collective geographical preferences. In Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on (pp. 312–317). IEEE. Retrieved from http://ieeexplore.ieee.org/abstract/document/5625119/

Calabrese, Francesco, Diao, M., Di Lorenzo, G., Ferreira, J., & Ratti, C. (2013). Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. Transportation Research Part C: Emerging Technologies, 26, 301–313. https://doi.org/10.1016/j.trc.2012.09.009

Cambridge Systematics. (2011). Chattanooga Cell Phone External OD Matrix Development-21 Process and Findings. Presented to Tennessee Model Users Group (TNMUG).

Chen, C., Bian, L. & Ma, J., (2014). From traces to trajectories: How well can we guess activity locations from mobile phone traces? Transportation Research Part C: Emerging Technologies, 46, 326–337.

Chen, C., Gong, H., Lawson, C., & Bialostozky, E. (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: Policy and Practice, 44(10), 830-840.

Chen, C., Ma, J., Susilo, Y., Liu, Y., & Wang, M. (2016). The promises of big data and small data for travel behavior (aka human mobility) analysis. Transportation research part C: emerging technologies, 68, 285-299.

Çolak, S., Alexander, L. P., Alvim, B. G., Mehndiratta, S. R., & González, M. C. (2015). Analyzing Cell Phone Location Data for Urban Travel. Transportation Research Record: Journal of the Transportation Research Board, 2526, 126–135. https://doi.org/10.3141/2526-14

Csáji, B. C., Browet, A., Traag, V. A., Delvenne, J.-C., Huens, E., Van Dooren, P., … Blondel, V. D. (2013). Exploring the mobility of mobile phone users. Physica A: Statistical Mechanics and Its Applications, 392(6), 1459–1473. https://doi.org/10.1016/j.physa.2012.11.040

Cui, Y., & Ge, S. S. (2003). Autonomous vehicle positioning with GPS in urban canyon environments. IEEE transactions on robotics and automation, 19(1), 15-25.

de Jong, R. and W. Mensonides (2003) Wearable GPS device as a data collection method for travel research,Working Paper, ITS-WP-03-02, University of Sydney, Institute of Transport Studies, Sydney.

Dingus, T. A., Klauer, S. G., Neale, V. L., Petersen, A., Lee, S. E., Sudweeks, J. D., ... & Bucher, C. (2006). The 100-car naturalistic driving study, Phase II-results of the 100-car field experiment (No. HS-810 593).

Dong, H., Wu, M., Ding, X., Chu, L., Jia, L., Qin, Y., & Zhou, X. (2015). Traffic zone division based on big data from mobile phone base stations. Transportation Research Part C: Emerging Technologies, 58, Part B, 278–291. https://doi.org/10.1016/j.trc.2015.06.007

Draijer, G., N. Kalfs and J. Perdok (2000) Global Positioning System as data collection method for travel research, Transportation Research Record, 1719, 147–153.

Du, J. and L. Aultman-Hall (2007) Increasing the accuracy of trip rate information from pas- sive multi-day GPS travel datasets: Automatic trip end identification issues, Transportation Research Part A: Policy and Practice, 41 (3) 220–232.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., & others. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd (Vol. 96, pp. 226–231). Retrieved from https://ocs.aaai.org/Papers/KDD/1996/KDD96-037.pdf

F. Calabrese, M. Diao, G.D. Lorenzo, J. Ferreira Jr., C. Ratti (2013) Understanding individual mobility patterns from urban sensing data: a mobile phone trace example Transp. Res. Part C, 26 (2013), pp. 301-313

Feng, T., & Timmermans, H. J. (2013). Transportation mode recognition using GPS and accelerometer data. Transportation Research Part C: Emerging Technologies, 37, 118-130.

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association, 97(458), 611–631.

Fussell, Gresham, & Smith. (n.d.). Origin Destination Analysis for Moore County, 27 NC.In For NCDOT & the Moore County Transportation Committee (MCTC). Parsons Brinckerhoff.

Gezici, S. (2008). A survey on wireless position estimation. Wireless Personal Communications, 44(3), 263–282.

Gong, H., Chen, C., Bialostozky*, E., and Lawson, C. (2012) A GPS/GIS Method for Travel Mode Detection in New York City. Computers, Environment, and Urban Systems, 36(2), 131-139.

Gong, L., Morikawa, T., Yamamoto, T., & Sato, H. (2014). Deriving personal trip data from GPS data: a literature review on the existing methodologies. Procedia-Social and Behavioral Sciences, 138, 557-565.

Gong, Lei, Hitomi Sato, Toshiyuki Yamamoto, and Tomio Miwa. 2015. “Identification of Activity Stop Locations in GPS Trajectories by Density-Based Clustering Method Combined with Support Vector Machines.” Journal of Modern Transportation 23 (3). Springer Berlin Heidelberg: 202–13. doi:10.1007/s40534-015-0079-x.

González, M. C., Hidalgo, C. A., & Barabási, A.-L. (2008). Understanding individual human mobility patterns. Nature, 453(7196), 779–782. https://doi.org/10.1038/nature06958

Gonzalez, P., Weinstein, J., Barbeau, S., Labrador, M., Winters, P., Georggi, N. L., & Perez, R. (2008, November). Automating mode detection using neural networks and assisted GPS data collected using GPS-enabled mobile phones. In 15th World congress on intelligent transportation systems (pp. 16-20).

GPS.gov, (2013), "How GPS Works" Poster. Retrieved from http://www.gps.gov/multimedia/poster/

Greater Buffalo-Niagara Transportation Survey 2002. (2003, April). Greater Buffalo-Niagara Regional Transportation Council.

Hao, P., Ban, X. J., Guo, D., & Ji, Q. (2014). Cycle-by-cycle intersection queue length distribution estimation using sample travel times. Transportation research part B: methodological, 68, 185-204.

Hao, P., Ban, X., & Yu, J. (2015). Kinematic equation-based vehicle queue location estimation method for signalized intersections using mobile sensor data. Journal of Intelligent Transportation Systems, 19(3), 256-272.

Hao, P., Ban, X., Bennett, K. P., Ji, Q., & Sun, Z. (2012). Signal timing estimation using sample intersection travel times. IEEE Transactions on Intelligent Transportation Systems, 13(2), 792-804.

Hard, E., Chigoy, B., Songchitruksa, P., Farnsworth, S., Borchardt, D., & Green, L. (2016). Synopsis of New Methods and Technologies to Collect Origin-Destination (OD) Data (No. FHWA-HEP-16-083).

Hariharan, R., & Toyama, K. (2004). Project Lachesis: Parsing and Modeling Location Histories. In M. J. Egenhofer, C. Freksa, & H. J. Miller (Eds.), Geographic Information Science (pp. 106–124).

Herrera, J. C., & Bayen, A. M. (2010). Incorporation of Lagrangian measurements in freeway traffic state estimation. Transportation Research Part B: Methodological, 44(4), 460-481.

Hunter, T., Herring, R., Abbeel, P., & Bayen, A. (2009). Path and travel time inference from GPS probe vehicle data. NIPS Analyzing Networks and Learning with Graphs, 12(1).

Huntsinger, L. F., & Donnelly, R. (2014). Reconciliation of regional travel model and passive device tracking data. In Transportation Research Board 9third Annual Meeting. Retrieved from https://trid.trb.org/view.aspx?id=1287620

Huntsinger, L. F., & Ward, K. (2015). Using mobile phone location data to develop external trip models. Transportation Research Record: Journal of the Transportation Research Board, (2499), 25–32.

Iovan, C., Olteanu-Raimond, A.-M., Couronné, T., & Smoreda, Z. (2013). Moving and Calling: Mobile Phone Data Quality Measurements and Spatiotemporal Uncertainty in Human Mobility Studies. In D. Vandenbroucke, B. Bucher, & J. Crompvoets (Eds.), Geographic Information

Science at the Heart of Europe (pp. 247–265). Cham: Springer International Publishing. Retrieved from http://dx.doi.org/10.1007/978-3-319-00615-4_14

Iqbal, M. S., Choudhury, C. F., Wang, P., & González, M. C. (2014). Development of origin–destination matrices using mobile phone call data. Transportation Research Part C: Emerging Technologies, 40, 63–74. https://doi.org/10.1016/j.trc.2014.01.002

Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M., Rowland, J., & Varshavsky, A. (2011). Identifying important places in people's lives from cellular network data. In International Conference on Pervasive Computing (pp. 133–151). Springer. Retrieved from http://link.springer.com/10.1007/978-3-642-21726-5_9

Jang, C. W., Juang, J. C., & Kung, F. C. (2000). Adaptive fault detection in real-time GPS positioning. IEE Proceedings-Radar, Sonar and Navigation, 147(5), 254-258.

Jiang, S., Ferreira, J., & Gonzalez, M. C. (2017). Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data: A Case Study of Singapore. IEEE Transactions on Big Data, PP(99), 1–1. https://doi.org/10.1109/TBDATA.2016.2631141

Jiang, Shan, Fiore, G. A., Yang, Y., Ferreira, J., Jr., Frazzoli, E., & González, M. C. (2013). A Review of Urban Computing for Mobile Phone Traces: Current Methods, Challenges and Opportunities. In Proceedings of the second ACM SIGKDD International Workshop on Urban Computing (p. 2:1–2:9). New York, NY, USA: ACM. https://doi.org/10.1145/2505821.2505828

Kami N, Enomoto N, Baba T, Yoshikawa T (2010) Algorithm for detecting significant locations from raw GPS data. Springer, Berlin, pp 221–235

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: analysis and implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7), 881–892.

Kochan, B., Bellemans, T., Janssens, D., & Wets, G. (2006). Dynamic activity-travel diary data collection using a GPS-enabled personal digital assistant. In Applications of Advanced Technology in Transportation (pp. 319-324).

Kunzmann, M., Daigler, V. (2013). 2010-2012 California Household Travel Survey Final Report. Retrieved from http://www.dot.ca.gov/hq/tpp/offices/omsp/statewide_travel_analysis/chts.html

Laasonen, K. (2005). Clustering and Prediction of Mobile User Routes from Cellular Data. In Knowledge Discovery in Databases: PKDD 2005 (pp. 569–576). Springer, Berlin, Heidelberg. https://doi.org/10.1007/11564126_59

Larijani, A. N., Olteanu-Raimond, A.-M., Perret, J., Brédif, M., & Ziemlicki, C. (2015). Investigating the mobile phone data to estimate the origin destination flow and analysis; case study: Paris region. Transportation Research Procedia, 6, 64–78.

Lee, J.-K., & Hou, J. C. (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 (pp. 85–96). New York, NY, USA: ACM. https://doi.org/10.1145/1132905.1132915

Leontiadis, I., Lima, A., Kwak, H., Stanojevic, R., Wetherall, D., & Papagiannaki, K. (2014). From Cells to Streets: Estimating Mobile Paths with Cellular-Side Data. In Proceedings of the 10th ACM International on Conference on Emerging Networking Experiments and Technologies (pp. 121–132). New York, NY, USA: ACM. https://doi.org/10.1145/2674005.2674982

Li, Z., Yu, L., Gao, Y., Wu, Y., Gong, D., & Song, G. (2016). Extraction method of temporal and spatial characteristics of residents' trips based on cellular signaling data, Transport Research, 1, 51-57

Lin, D.-B., & Juang, R.-T. (2005). Mobile location estimation based on differences of signal attenuations for GSM systems. IEEE Transactions on Vehicular Technology, 54(4), 1447–1454.

Liu, H., Danczyk, A., Brewer, R., & Starr, R. (2008). Evaluation of Cell Phone Traffic Data in Minnesota. Transportation Research Record: Journal of the Transportation Research Board, 2086, 1–7. https://doi.org/10.3141/2086-01

Liu, L., Biderman, A., & Ratti, C. (2009, June). Urban mobility landscape: Real time monitoring of urban mobility patterns. In Proceedings of the 11th International Conference on Computers in Urban Planning and Urban Management (pp. 1-16).

Liu, Y., Wang, F., Xiao, Y., & Gao, S. (2012). Urban land uses and traffic 'source-sink areas': Evidence from GPS-enabled taxi data in Shanghai. Landscape and Urban Planning, 106(1), 73-87.

Lu, S., Fang, Z., Zhang, X., Shaw, S.-L., Yin, L., Zhao, Z., & Yang, X. (2017). Understanding the Representativeness of Mobile Phone Location Data in Characterizing Human Mobility Indicators. ISPRS International Journal of Geo-Information, 6(1), 7. https://doi.org/10.3390/ijgi6010007

Ma, J., Li, H., Yuan, F., & Bauer, T. (2013). Deriving Operational Origin-Destination Matrices From Large Scale Mobile Phone Data. International Journal of Transportation Science and Technology, 2(3), 183–204. https://doi.org/10.1260/2046-0430.2.3.183

Maerivoet, S., & Logghe, S. (2007). Validation of travel times based on cellular floating vehicle data. In Proceedings of the 6th European Congress and Exhibition on Intelligent Transport Systems and Services, Aalborg, Denmark (pp. 18–20). Retrieved from https://www.researchgate.net/profile/Steven_Logghe/publication/228922403_Validation_of_travel_times_based_on_cellular_floating_vehicle_data/links/004635257e1eb33f77000000.pdf

Milone. (2015). Evaluation of Cellular Origin-Destinaiton Data as a Basis for 34 Forecasting Non-Resident Travel (PowerPoint Presentation) (In 15th TRB National 35 Transporation Planning Applications Conference, Atlantic City, New Jersey).

Mizuno K, Kanamori R, Sano S, Nakajima S, Ito T (2013) Identifying move and stop in GPS data with Support Vector Machines. conference of infrastructure planning and management (CD-ROM), JSCE

Mohr, M., Edwards, C., & McCarthy, B. (2008). A study of LBS accuracy in the UK and a novel approach to inferring the positioning technology employed. Computer Communications, 31(6), 1148–1159.

Murakami, E., & Wagner, D. P. (1999). Can using global positioning system (GPS) improve trip reporting?. Transportation research part c: emerging technologies, 7(2), 149-165.

Murakami, E., Wagner, D. P., & Neumeister, D. M. (2004, July). Using global positioning systems and personal digital assistants for personal travel surveys in the United States. In International Conference on Transport Survey Quality and Innovation.

Nitsche, P., Widhalm, P., Breuss, S., Brändle, N., & Maurer, P. (2014). Supporting large-scale travel surveys with smartphones–A practical approach. Transportation Research Part C: Emerging Technologies, 43, 212–221.

Palma AT, Bogorny V, Kuijpers B, Alvares LO (2008) A clustering-based approach for discovering interesting places in trajectories. Proceedings of the 2008 ACM symposium on applied computing. ACM. pp 863–868

Pearson, D. (2001) Global Positioning System (GPS) and travel surveys: Results from the 1997 Austin household survey, paper presented at 8th Conference on the Application of Transportation Planning Methods, Corpus Christi, April 2001.

Peng, C., Jin, X., Wong, K. C., Shi, M., & Liò, P. (2012). Collective human mobility pattern from taxi trips in urban area. PloS one, 7(4), e34487.

Phatak, M., Chansarkar, M., & Kohli, S. (1999). Position fix from three GPS satellites and altitude: a direct method. IEEE Transactions on Aerospace and Electronic Systems, 35(1), 350-354.

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

Qi, L., Qiao, Y., Abdesslem, F. B., Ma, Z., & Yang, J. (2016). Oscillation Resolution for Massive Cell Phone Traffic Data. In Proceedings of the First Workshop on Mobile Data (pp. 25–30). New York, NY, USA: ACM. https://doi.org/10.1145/2935755.2935759

Qu, Y., Gong, H., & Wang, P. (2015). Transportation Mode Split with Mobile Phone Data. In 2015 IEEE 18th International Conference on Intelligent Transportation Systems (pp. 285–289). https://doi.org/10.1109/ITSC.2015.56

Rakha, H., & Van Aerde, M. (1995, March). Accuracy of vehicle-probe estimates of link-travel time and instantaneous speed. In Proceedings of the Annual Meeting of ITS America, Washington DC (pp. 385-92).

Ratti, C., Frenchman, D., Pulselli, R. M., & Williams, S. (2006). Mobile Landscapes: Using Location Data from Cell Phones for Urban Analysis. Environment and Planning B: Planning and Design, 33(5), 727–748. https://doi.org/10.1068/b32047

Ray, J. K., Cannon, M. E., & Fenton, P. (2001). GPS code and carrier multipath mitigation using a multiantenna system. IEEE Transactions on Aerospace and Electronic Systems, 37(1), 183-195.

Reddy, S., Mun, M., Burke, J., Estrin, D., Hansen, M., & Srivastava, M. (2010). Using mobile phones to determine transportation modes. ACM Transactions on Sensor Networks (TOSN), 6(2), 13.

RSG. (2015). Big Data in Transportation Planning: User Perspectives, Challenges and Successes. Idaho.

Sanwal, K. K., & Walrand, J. (1995). Vehicles as probes. California Partners for Advanced Transit and Highways (PATH).

Schüssler, N., & Axhausen, K. W. (2009). Processing GPS raw data without additional information. Transportation Research Record, 2105, 28-36.

Schwarzenegger, A., Bonner, D. E., Kempton, W., & Copp, R. (2008). State highway congestion monitoring program (HICOMP), annual data compilation. Technical report, Caltrans, Sacramento, CA.

Shoval, N. (2008). Tracking technologies and urban analysis. Cities, 25(1), 21-28.

Skyhook Wireless. (2008). WiFi Positioning System: Accuracy, Availability and Time to Fix Performance (Wireless Technical White Paper). Boston, MA.

Smith, B. L., Pack, M. L., Lovell, D. J., & Sermons, M. W. (2001). Transportation management applications of anonymous mobile call sampling. In ITS America 11th Annual Meeting and Exposition, ITS: Connecting the Americas. Retrieved from https://trid.trb.org/view.aspx?id=693557

Song, C., Koren, T., Wang, P., & Barabási, A.-L. (2010). Modelling the scaling properties of human mobility. Nature Physics, 6(10), 818–823. https://doi.org/10.1038/nphys1760

Stenneth, L., Wolfson, O., Yu, P. S., & Xu, B. (2011, November). Transportation mode detection using mobile phones and GIS information. In Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (pp. 54-63). ACM.

Stopher, P. R, Jiang, Q. & FitzGerald, C. (2005). Processing GPS data from travel surveys. Proceedings of second Int. Colloquium on the Behavioral Foundations of Integrated Land-use and Transportation Models: Frameworks, Models and Applications. June 2005. Toronto, Canada.

Stopher, P. R., & Greaves, S. P. (2007). Household travel surveys: Where are we going?. Transportation Research Part A: Policy and Practice, 41(5), 367-381.

Stopher, P. R., Prasad, C., & Zhang, J. (2010). Can GPS Replace Conventional Travel Surveys?: Some Findings. Institute of Transport and Logistics Studies.

Stopher, P., Bullock, P. & Jiang, Q. (2002). GPS, GIS and personal travel surveys: an exercise in visualisation, 25th Australasian Transport Research Forum Incorporating the BTRE Transport Policy Colloquium, October 2002, Canberra.

Stopher, P., Clifford, E., Zhang, J. & FitzGerald, C. (2008a). Deducing mode and purpose from GPS data. Working paper ITLS-WP-08-06. Institute of Transport and Logistic Studies, the Australian Key Center in Transport and Logistic Management, the University of Sydney.

Stopher, P., FitzGerald, C., & Zhang, J. (2008b). Search for a global positioning system device to measure person travel. Transportation Research Part C: Emerging Technologies, 16(3), 350-369.

Streetlight Insight, (2012). Retrieved from https://www.streetlightdata.com/transportation-planning-product

Su, W., Lee, S. J., & Gerla, M. (2000). Mobility prediction in wireless networks. In MILCOM 2000. 2first Century Military Communications Conference Proceedings (Vol. 1, pp. 491-495). IEEE.

Sun, Z., Hao, P., Ban, X. J., & Yang, D. (2015). Trajectory-based vehicle energy/emissions estimation for signalized arterials using mobile sensing data. Transportation Research Part D: Transport and Environment, 34, 27-40.

Syed, S., & Cannon, M. E. (2004, January). Fuzzy logic-based map matching algorithm for vehicle navigation system in urban canyons. In ION National Technical Meeting, San Diego, CA (Vol. 1, pp. 26-28).

Tao, S., Manolopoulos, V., Rodriguez Duenas, S., & Rusu, A. (2012). Real-time urban traffic state estimation with A-GPS mobile phones as probes. Journal of Transportation Technologies, 2(1), 22-31.

Tettamanti, T., & Varga, I. (2014). Mobile phone location area based traffic flow estimation in urban road traffic. Columbia International Publishing, Advances in Civil and Environmental Engineering, 1(1), 1–15.

Tettamanti, T., Demeter, H., & Varga, I. (2012). Route choice estimation based on cellular signaling data. Acta Polytechnica Hungarica, 9(4), 207–220.

Thiessenhusen, K. U., Schafer, R. P., & Lang, T. (2003). Traffic data from cell phones: a comparison with loops and probe vehicle data. Institute of Transport Research German Aerospace Center, Germany.

Thiessenhusen, K.-U., Schäfer, R.-P., & Lang, T. (2006). Traffic Data from Cell Phones-a Comparison with Loops and Floating Car Data. In PROCEEDINGS OF THE 13th ITS WORLD CONGRESS, LONDON, 8-12 OCTOBER 2006.

Toole, J. L., Herrera-Yaqüe, C., Schneider, C. M., & González, M. C. (2015). Coupling human mobility and social ties. Journal of The Royal Society Interface, 12(105), 20141128. https://doi.org/10.1098/rsif.2014.1128

Tran LH, Nguyen QVH, Do NH, Yan Z (2011) Robust and hierarchical stop discovery in sparse and diverse trajectories (No. EPFL-REPORT-175473)

Van Diggelen, F., & Enge, P. (2015, September). The world's first GPS MOOC and worldwide laboratory using smartphones. In Proceedings of the 28th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS+ 2015) (pp. 361-369).

Wagner, D. P. (1997) Lexington area travel data collection test: GPS for personal travel surveys, Final Report, Office of Highway Policy Information and Office of Technology Applications, Federal Highway Administration, Battelle Transport Division, Columbus, September 1997.

Wagner, D.P., E. Murakami, and D.M. Neumeister (1968), "Global Positioning System for Personal Travel Surveys", Federal Highway Administration, Washington, DC.

Walsh, D., Capaccio, S., Lowe, D., Daly, P., Shardlow, P., & Johnston, G. (1997). Real time differential GPS and GLONASS vehicle positioning in urban areas. Space communications, 14(4), 203-217.

Wang, F., & Chen, C. (2017). On data processing required to derive mobility patterns from passively-generated mobile phone data. (In Press) Transportation Research Part C.

Wang, H., Calabrese, F., Lorenzo, G. D., & Ratti, C. (2010). Transportation mode inference from anonymized and aggregated mobile phone call detail records. In 2010 13th International IEEE Conference on Intelligent Transportation Systems (ITSC) (pp. 318–323). https://doi.org/10.1109/ITSC.2010.5625188

Wang, M., 2014. Understanding Activity Location Choice with Mobile Phone Data, Ph.D. Dissertation, Civil and Environmental Engineering, University of Washington, Seattle.

Wang, P., Hunter, T., Bayen, A. M., Schechtner, K., & González, M. C. (2012). Understanding road usage patterns in urban areas. Scientific Reports, 2, 1001.

Weiss, A. J. (2003). On the accuracy of a cellular location system based on RSS measurements. IEEE Transactions on Vehicular Technology, 52(6), 1508–1518.

Welbourne, E., Lester, J., LaMarca, A., & Borriello, G. (2005). Mobile context inference using low-cost sensors. Location-and Context-Awareness, 95–127.

Widhalm, P., Nitsche, P., & Brändie, N. (2012, November). Transport mode detection with realistic smartphone sensor data. In Pattern Recognition (ICPR), 2012 2first International Conference on (pp. 573-576). IEEE.

Widhalm, P., Yang, Y., Ulm, M., Athavale, S., & González, M. C. (2015). Discovering urban activity patterns in cell phone data. Transportation, 42(4), 597–623. https://doi.org/10.1007/s11116-015-9598-x

Wolf, J. (2006). Applications of new technologies in travel surveys. In Travel survey methods: Quality and future directions (pp. 531-544). Emerald Group Publishing Limited.

Wolf, J., Guensler, R., &Bachman, W. (2001). Elimination of the Travel Diary: An Experiment to Derive Trip Purpose From GPS Travel Data. Proceedings of the 80th Annual Meeting of the Transportation Research Board, January 2001, Washington D.C.

Work, D. B., Tossavainen, O. P., Blandin, S., Bayen, A. M., Iwuchukwu, T., & Tracton, K. (2008, December). An ensemble Kalman filtering approach to highway traffic estimation using GPS enabled mobile devices. In Decision and Control, 2008. CDC 2008. 47th IEEE Conference on (pp. 5062-5068). IEEE.

Wu, W., Wang, Y., Gomes, J. B., Anh, D. T., Antonatos, S., Xue, M., Nash, A. S. (2014). Oscillation Resolution for Mobile Phone Cellular Tower Data to Enable Mobility Modelling. In 2014 IEEE 15th International Conference on Mobile Data Management (Vol. 1, pp. 321–328). https://doi.org/10.1109/MDM.2014.46

Xiao, G., Juan, Z., & Zhang, C. (2015). Travel mode detection based on GPS track data and Bayesian networks. Computers, environment and urban systems, 54, 14-22.

Xu, C., Ji, M., Chen, W., & Zhang, Z. (2010, August). Identifying travel mode from GPS trajectories through fuzzy pattern recognition. In Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on (Vol. 2, pp. 889-893). IEEE.

Yang, X., Sun, Z., Ban, X., & Holguín-Veras, J. (2014). Urban Freight Delivery Stop Identification with GPS Data. Transportation Research Record: Journal of the Transportation Research Board, 2411, 55–61. https://doi.org/10.3141/2411-07

Ye, X., Konduri, K., Pendyala, R. M., Sana, B., & Waddell, P. (2009). A methodology to match distributions of both household and person attributes in the generation of synthetic populations. In 88th Annual Meeting of the Transportation Research Board, Washington, DC. Retrieved from http://www.scag.ca.gov/Documents/PopulationSynthesizerPaper_TRB.pdf

Ye, Y., Zheng, Y., Chen, Y., Feng, J., & Xie, X. (2009). Mining individual life pattern based on location history. The 10th International Conference on Mobile Data Management: Systems, Services and Middleware, IEEE. DOI: 10.1109/MDM.2009.11.

Yim, Y. B. Y., & Cayford, R. (2001). Investigation of Vehicles as Probes Using Global Positioning System and Cellular Phone Tracking: Field Operational Test. California Partners for Advanced Transit and Highways (PATH). Retrieved from http://escholarship.org/uc/item/0378c1wc

Yin, M., Sheehan, M., Feygin, S., Paiement, J.-F., & Pozdnoukhov, A. (2017). A generative model of urban activities from cellular Data. IEEE Transactions on Intelligent Transportation Systems. Retrieved from http://ieeexplore.ieee.org/abstract/document/7932990/

Yuan, Y., Raubal, M., & Liu, Y. (2012). Correlating mobile phone usage and travel behavior – A case study of Harbin, China. Computers, Environment and Urban Systems, 36(2), 118–130. https://doi.org/10.1016/j.compenvurbsys.2011.07.003

Zandbergen, P. A. (2009). Accuracy of iPhone Locations: A Comparison of Assisted GPS, WiFi and Cellular Positioning. Transactions in GIS, 13, 5–25. https://doi.org/10.1111/j.1467-9671.2009.01152.x

Zhan, X., Hasan, S., Ukkusuri, S. V., & Kamga, C. (2013). Urban link travel time estimation using large-scale taxi data with partial information. Transportation Research Part C: Emerging Technologies, 33, 37-49.

Zhang, L., Dalyot, S., Eggert, D., & Sester, M. (2011). Multi-stage approach to travel-mode segmentation and classification of GPS traces. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences:[Geospatial Data Infrastructure: From Data Acquisition And Updating To Smarter Services] 38-4 (2011), Nr. W25, 38(W25), 87-93.

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

Zheng, Y., Liu, L., Wang, L., & Xie, X. (2008, April). Learning transportation mode from raw GPS data for geographic applications on the web. In Proceedings of the 17th international conference on World Wide Web (pp. 247-256). ACM.

Zhou, C., Jia, H., Juan, Z., Fu, X., & Xiao, G. (2016). A Data-Driven Method for Trip Ends Identification Using Large-Scale Smartphone-Based GPS Tracking Data, 1–15.

Zimmermann M, Kirste T, Spiliopoulou M (2009) Finding stops in error-prone trajectories of moving objects with time-based clustering. Intelligent interactive assistance and mobile multi-media computing. Springer, Berlin, pp 275–286

NOTICE

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The United State Government assumes no liability for its contents or use thereof.

The United States Government does not endorse manufacturers or products. Trade names appear in the document only because they are essential to the content of the report.

The opinions expressed in this report belong to the authors and do not constitute an endorsement or recommendation by FHWA.

This report is being distributed through the Travel Model Improvement Program (TMIP).

U.S. Department of Transportation
Federal Highway Administration
Office of Planning, Environment, and Realty
1200 New Jersey Avenue, SE
Washington, DC 20590

February 2019

FHWA-HEP-19-027

U.S.Department of Transportation
Federal Highway Administration