

Promises of Data from Emerging Technologies for Transportation Applications: Puget Sound Region Case Study

WA-RD 892.1

Xuegang (Jeff) Ban
Feilong Wang
Yiran Zhang

Cynthia Chen
Jingxing Wang

December 2018



**Washington State
Department of Transportation**

Office of Research & Library Services

WSDOT Research Report

Promises of Data from Emerging Technologies for Transportation Applications: Puget Sound Region Case Study

DECEMBER 2018



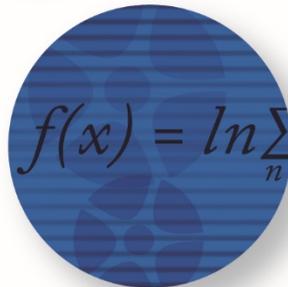
Peer Reviews



Data



Tools



$f(x) = \ln \sum_n$



Webinars



Analysis



U.S. Department of Transportation
Federal Highway Administration



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. WA-RD 892.1	2. Government Accession No.	3. Recipient's Catalog No.	
Promises of Data from Emerging Technologies for Transportation Applications: Puget Sound Region Case Study		5. Report Date December 2018	
		6. Performing Organization Code	
7. Authors Xuegang (Jeff) Ban, Cynthia Chen, Feilong Wang, Jingxing Wang, Yiran Zhang		8. Performing Organization Report No.	
9. Performing Organization Name and Address Department of Civil and Environmental Engineering University of Washington		10. Work Unit No. (TR AIS)	
		11. Contract or Grant No. FY18004 / T1461-53	
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 January 2018 to December 2018	
		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 With the explosion of the number of studies using big, passively-generated data for transportation analysis, this study focuses on understanding the properties of such data and how these properties affect our ability in deriving trip-related characteristics. Two big datasets were analyzed: a mobile phone data generated primarily on phone calls with locations identified through cellular triangulation and an app-based data generated primarily on app usage with locations identified through a mix of positioning technologies including GPS and cellular triangulation. Both datasets were compared against their household travel survey counterparts. It is shown that the two datasets, generated through different positioning technologies and usage mechanisms clearly have different spatial and temporal characteristics, which then affect trip related attributes such as trip rates and OD patterns. Implications in planning applications and future work are discussed.			
17. Key Words Big Data, Mobile Phone Data, App-Based Data, Travel Surveys, Travel Patterns, Origin Destination Demand Matrices		18. Distribution Statement No restrictions.	
19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 120	22. Price 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)

Promises of Data from Emerging Technologies for Transportation Applications: Puget Sound Region Study

December 2018

Federal Highway Administration

Table of Contents

Notice	2
Quality Assurance Statement	2
Table of Contents	v
List of Figures	vii
List of Tables	ix
Lists of Abbreviations and Symbols	x
Abbreviations.....	x
Symbols	x
1.0 Executive Summary	13
1.1 Research Questions and Findings	13
1.2 Recommendations.....	16
2.0 Introduction	19
2.1 Disclaimer.....	19
2.2 Acknowledgments	19
2.3 From Travel Survey Data to Passively Generated Emerging Data.....	19
3.0 App-Based Data	21
3.1 General Description (Zeroth Order)	22
3.2 First-Order Properties	33
3.3 Second-Order Properties	40
3.4 May 9th Data Shift	48
3.5 Summary	55
3.6 Discussion	63
4.0 Other Emerging Data Sources and Applications	67
4.1 Data from Connected Vehicles	67
4.2 Data from Automated Vehicles	75
4.3 Data from New Shared Mobility Services.....	77
5.0 Development of A Data Fusion Framework	85
5.1 Goal and Objectives of Data Fusion	88
5.2 Principles of Developing Data Fusion Methods.....	88
6.0 References	90
Appendixes	93

A.1	Appendix A—Extracting Trips from the App-based Data	93
A.2	Appendix B—OD Estimation Method for App-Based Data	97
A.3	Appendix C—Home Distribution of Users Observed Every Day	98
A.4	Appendix D—CV data	103
A.5	Appendix E—Shared Mobility Data	110
	Appendix References	116

List of Figures

Figure 1. Map. Spatial distribution of observations	23
Figure 2. Map. A zoom-in view of the central PSR during the evening peak of a typical weekday.....	24
Figure 3. Graph. Time interval distribution between two consecutive observations	25
Figure 4. Graph. Distribution of location accuracy	26
Figure 5. Graph. Cumulative distribution of location accuracy	26
Figure 6. Graph. Distribution of observations within a day (the Week of April 17th).....	27
Figure 7. Graph. Weekly pattern of observations (Sundays are in open box)	28
Figure 8. Graph. Percentage of trajectories with their locations revealed at a time of the day, comparing different days in a week.	29
Figure 9. Graph. Percentage of trajectories with their locations revealed at different times of a holiday.	29
Figure 10. Graph. Inter-day sparsity (distribution of life span of unique IDs).....	30
Figure 11. Graph. Inter-day sparsity (distribution of number of days observed).....	31
Figure 12. Graph. Distribution of temporal resolution of all (daily) trajectories.	32
Figure 13. Map. Comparison between inferred home locations from the app-based data and the population from the census. (a) Inferred home location density and (b) Census population density.	34
Figure 14. Graph. Correlation between inferred home locations and census population.	35
Figure 15. Graph. Distribution of scaling factor.....	36
Figure 16. Illustration. Activity duration and a demonstration of the biased estimation.	36
Figure 17. Graph. Activity duration observed from PSRC survey and app-based data	37
Figure 18. Graph. Spatial distribution of trip ends on a weekday morning.	38
Figure 19. Graph. Spatial distribution illustrating where more trip ends are observed on weekdays than that on weekends (in TAZ).....	39
Figure 20. Graph. Distribution of trip rates.....	40
Figure 21. Graph. Weekly trip rate pattern	41
Figure 22. Graph. Departure time distribution	42
Figure 23. Graph. Distribution of travel distance.....	43
Figure 24. Graph. Cumulative distribution of travel distance.....	43
Figure 25. Graph. Distribution of travel times	44
Figure 26. Graph. Cumulative distribution of travel times	45
Figure 27. Map. Spatial distribution of trip origins. (a) Estimated results from app-based data, (b) SoundCast results.....	46
Figure 28. Graph. Correlations between estimated trip origins and MPO trip origins.....	47
Figure 29. Graph. Correlations between estimated OD demands and PSRC OD demands	47
Figure 30. Graph. Evolution of daily number of unique IDs (zeroth order)	48
Figure 31. Graph. Evolution of daily number of observations per ID (zeroth.....	49
Figure 32. Graph. Evolution of location accuracy (zeroth order).....	50
Figure 33. Graph. Evolution of time interval (zeroth order)	51
Figure 34. Graph. Evolution of temporal sparsity (zeroth order)	51

Figure 35. Graph. Comparison of activity duration (first^t order) before and after May 9th52

Figure 36. Graph. Comparison of trip rate (second order) before and after May 9th53

Figure 37. Graph. Comparison of departure time distributions (second order) before and after May 9th54

Figure 38. Graph. Comparison of cumulative distributions of trip length (second order) before and after May 9th54

Figure 39. Graph. Comparison of cumulative distributions of travel time (second order) before and after May 9th55

Figure 40. Graph. Fraction of IDs with their locations revealed at time of a day (both big data sets)59

Figure 41. Graph. Fraction of observations within the week (both big data sets)60

Figure 42. Graph. Cumulative distribution of data location accuracy (both big data sets)61

Figure 43. Graph. Cumulative distribution of travel times (four data sets).....63

Figure 44 Graph. Layered architecture of dedicated short-range communications (DSRC) [12].....68

Figure 45 Graph. Process to select applications (CV102: Participant Workbook Sept 2015)74

Figure 46. Lyft API78

Figure 47. Uber Movement user interface79

Figure 48 Graph. Integration of big data and small data.....85

Figure 49. Graph. Illustration of variable definitions for stay identification.....93

Figure 50. Illustration. Illustration of identifying stays from the GPS data set. (a) Raw GPS trajectories of two days; (b) Processed trajectories with identified stays; (c) Processed trajectories with a common stay being identified.95

Figure 51. Illustration. Demonstration of the spatiotemporal relationship. (a) Temporally separate and spatially contiguous; (b) Temporally contained (c) Temporally intersected (cutting off t_{a3} turns a into a' which is temporally separate with b).96

Figure 52. Illustration. Definition of spatially contiguous.97

Figure 53. Map. Comparison between home locations of IDs observed everyday (8,758 IDs) and the population from the census. (a) Home location density of IDs observed every day and (b) Census population density.99

Figure 54. Graph. Correlation between home locations of IDs observed every day and census population.100

Figure 55. Map. Comparison between home locations of users with life span of 63 days (41,640 IDs) and the population from the census. (a) Home location density of IDs with life span of 63 days and (b) Census population density.101

Figure 56. Comparison of home locations of IDs with life span of 63 days (41,640 IDs).....102

Figure 57. Sample accident report of AV (1)107

Figure 58. Sample accident report of AV (2)108

Figure 59. Sample accident report of AV (3)109

List of Tables

Table 1. An overview of the analysis framework.....	21
Table 2. A synthetic sample of app-based data.....	22
Table 3. Summary of datasets	56
Table 4. Pros/cons of different data.....	57
Table 5. Messages defined in SAE J2735_201603 (Dedicated Short Range Communications (DSRC) Message Set Dictionary, 2016).....	69
Table 6. Data frame of core Basic Safety Message (BSM) (Wyoming CV Pilot Basic Safety Message One Day Sample - (Data.gov)).....	70
Table 7. Data format of SPaT Messages (“Data.gov,”).....	72
Table 8. The applications selected by the USDOT to utilize in a CV Pilot Program	73
Table 9. Levels of driving automation (NHSTA & SAE)	76
Table 10. Autonomous vehicle accident data format [8].....	77
Table 11. Data format (All DATA from Uber Movement).....	80
Table 12. Data format (Trajectory Data).....	81
Table 13. Data format (Order Data).....	81
Table 14. Data format (Bike Sharing Demand Kaggle).....	82
Table 15. Characteristics of different data sets.....	87
Table 16. Origin to All Destination.....	110
Table 17. Daily time series (evening selected)	111
Table 18. Chart data (day of week, from 1/1/2018-1/31/2018).....	112
Table 19. ALL DATA (month aggregate)	113
Table 20. Raw data of trajectory data from DiDi (city of Xi’an, China, 2016/10/30).....	114
Table 21. Raw data of order data from DiDi (city of Chengdu, China)	114
Table 22. Bike sharing demand.....	115

Lists of Abbreviations and Symbols

Abbreviations

AGPS	Assisted GPS
API	Application Programming Interface
AV	Automated Vehicles
BSM	Basic Safety Message
CAVs	Connected and Automated Vehicles
CVs	Connect Vehicles
DCI	Divide, Conquer and Integrate
DOT	Department of Transportation
DSRC	Dedicated Short-Range Communications
FCW	Forward Collision Warning
GIS	Geographic Information System
GPS	Global Positioning System
IMA	Intersection Movement Assistance
JSON	JavaScript Object Notation
MPOs	Metropolitan Planning Organizations
OD	Origin-Destination
PSR	Puget Sound Region
PSRC	Puget Sound Regional Council
RLVW	Red Light Violation Warning
SDKs	Software Development Kits
SPaT	Signal Phase and Timing
TAZ	Traffic Analysis Zone
TNC	Transportation Network Company
V2I	Vehicle to Infrastructure
V2P	Vehicle to Pedestrian
V2V	Vehicle to Vehicle
V2X	Vehicle to Everything

Symbols

$Trajectory_{id}$	the sequence of observations of user i on day d
Φ_{id}	temporal resolution, the number of time slots in which this anonymous user is observed at least once
\hat{t}_{arr}^i	the observed arrival time of activity i
t_{arr}^i	the actual arrival time of activity i
\hat{t}_{dep}^i	the observed departure time of activity i
t_{dep}^i	the actual departure time of activity i
Δl_{roam}	signal roaming distance
Δt_{dur}	stay duration
$scc(tm, tm+1, \dots, tn; lngcc, latcc; rcc)$	sequence of observations representing a stay
$(lngcc, latcc)$	the centroid of the cluster where sc belongs
Rcc	the radii of the cluster, the longest distance from the centroid to any stays in the cluster
Tc	duration threshold
Rc	spatial threshold
D_{ab}	the distance between centroids of two stays (a, b)

α_i	the scaling factor of <i>TAZ i</i>
P_i	the population of <i>TAZ i</i>
r_i	the number of residents of <i>TAZ i</i>
$Trip_{(i,a,b)}$	the number of observed trips between TAZ pair (<i>a, b</i>) and generated by the user associated with <i>TAZ i</i>
$OD_{(a,b)}$	the OD demand between TAZ pair (<i>a, b</i>)

1.0 Executive Summary

The prevalence of mobile devices and the associated location positioning technologies that are needed to enable network connection at any time and any place have led to an explosion of studies that have used the resulting data (often big) for understanding travel patterns and to potentially guide policies. Such data are typically termed “passively solicited data.” These are different from the actively solicited data that result from a rigorously designed, probabilistic sampling process with a known target population. Instead, they are the secondary product of primary activities such as billing or operations (e.g., facilitating phone calls or use of mobile apps). Hence **the passively solicited data underlying the data generation process are often unknown, uncontrolled, and non-probabilistic**, raising questions regarding representativeness, accuracy, and stability of the estimates derived from such data.

One of the major objectives of this research is to demonstrate the importance of knowing your big data before any application, especially in the context of generating origin and destination patterns. In contrast to the vast majority of the studies that have used big data to derive trip-related statistics (e.g., trip rates), this study focused on understanding passively solicited data by developing a three-order analysis framework in which three groups of statistics were calculated. These statistics relate to the data themselves (zeroth order), single locations or trip ends (first order), and a pair of locations or trips (second order). Two types of passive data were analyzed: mobile phone data triggered primarily from phone calls with locations identified through cellular triangulation, and app-based data generated primarily from apps usage, with locations identified through a mix of positioning technologies including GPS and cellular triangulation. These two types of data reflect the evolution of technologies being used to generate such data. Within the two-month study period of the app-based data, an additional technology change in capturing the data occurred, resulting in a roughly 33 percent increase in the number of observations per device. This offered another opportunity to understand how the stability of the first- and second-order properties (those relating to trip ends and trips) may be affected by changes in technologies used to generate the data. More specifically, the study sought to answer six specific questions (see Section 1.1 below) relating to the data themselves, their implications for deriving trip-related characteristics such as trip rates and origin-destination (OD) patterns, and how we should leverage different types of data—big and small.

1.1 Research Questions and Findings

1. What analysis framework and associated metrics can be used to capture various properties of the passively solicited data?

As briefly noted earlier, a three-order analysis framework was proposed to capture the properties associated with the data themselves (zeroth order), single locations or trip ends (first order), and a pair of locations or trips (second order). This framework captures all related characteristics in a complete and logical way (for details, please refer to Table 1).

2. What is our current understanding of passively solicited data through the proposed three-order analysis framework?

The two most important zero-order properties of the data relate to how well the device is identified in spatial and temporal spaces. The spatial dimension is captured by “locational accuracy,” referring to the uncertainty involved in the positioning of the device (the smaller the better); its temporal dimension is captured by “temporal sparsity,” referring to the spread of the observations over a day (the more spread the better). **From the mobile phone data to the app-based data, we observed a significant improvement in locational accuracy (the 85th percentiles for the mobile phone and apps data were 700 and 100 meters, respectively, as shown in Figure 42) because of the prevalent usage of GPS-based navigation apps.** On the temporal dimension, improvement also occurred, although at a smaller magnitude; in comparison to locations in the mobile phone data, about 20 percent more location trajectories in the app-based data were revealed during the time period of 00:00 to 06:00 AM (see Figure 40). Additionally, two peaks (morning and afternoon) emerged in the apps data, as opposed to a single afternoon peak in the mobile phone data, although the two peaks were delayed in comparison to the traffic peaks in the region.

The zeroth-order properties have important implications for the first- and second-order characteristics. In particular, **the temporal sparsity and location accuracy noted above were found to directly impact the accuracy of the identified activity locations**, thus determining the activity (i.e., first order) and trip (i.e., second order) related characteristics. The technological improvement in capturing locations and the more dispersed app usage throughout a day (as compared to phone calls) allowed better capturing of home census tracts¹ and trip rates. The change from mobile phone data to app-based data resulted in data that more closely resembled the household travel survey data for trip rates (3.23 from apps data vs 4.40 from PSRC travel survey data, compared to 1.78 from mobile phone data vs 3.89 from Buffalo travel survey data). And correlation with population density at the census tract level increased from 0.43 to 0.91. However, the verdict on other statistics, such as activity duration, departure time, and OD patterns, was much less clear. This indicates that more in-depth analysis among the zeroth- and first- and second-order statistics needs to be done to gain a systematic understanding about how the data properties affect our ability to derive trip-related characteristics.

3. As the underlying data generation process changes, leading to changes in spatial and temporal properties as well as changes in trip-related metrics, how shall we interpret the resulting changes?

Clearly, improvement in data quality, both in terms of locational accuracy (Figure 4) and temporal sparsity (Figure 12), benefited more accurate calculation of metrics such as home census tracts and trip rates. But questions still remain: **for frequency of observations, is more always better? Or is there a threshold after which the bias of under-estimation becomes that of over-estimation?** Within the apps data, we also observed that when there was a 33 percent increase in the number of observations per device and consequently an improvement in temporal sparsity, the average trip rate consequently increased from 3.11 to 3.47, edging closer to the 4.4 from the household travel survey. However, the difference was not apparent for other

¹ For privacy issues, inferred home locations are presented at census tract level throughout the report.

metrics such as activity duration, departure time, or trip length. We therefore conclude the following:

- 1) When temporal sparsity is relatively low (which is the case for both mobile phone data and app-based data) and locational accuracy is low (which is the case for mobile phone data), improvement in both will likely move metrics closer to the ground truth.
- 2) However, as temporal sparsity continues to increase, the marginal benefit decreases. In fact, beyond a certain threshold, we suspect the positive benefit may even become negative, although this will require future research.
- 3) Improvement in different metrics may vary, as shown by trip rate, activity duration, departure time, and trip length.

4. Can we be more proactive in estimating trip-related metrics as the technologies and other circumstances underlying the big data generation process change over time?

The technologies used to generate the big data will inevitably change. Consequently, it will be worthwhile to ask whether we can be ahead of changes by being able to predict the consequences of the changes, i.e., **how will a sudden increase in locational accuracy and temporal sparsity affect our estimates of trips?** As shown in our answers to questions 2 and 3 above, this study demonstrated the inherent relationship between zeroth-order properties of the data themselves and the first- and second-order characteristics of trip ends and trips. More in-depth analysis is required to gain a systematic understanding of the nature of these relationships, which would certainly empower us the predictive capability.

5. How do we deal with the issue that big data lack ground truth?

As noted by Chen et al. (2014, 2016), **because of the uncontrolled data generation process associated with big data, validation of the inferred statistics from the data is critically important. And yet, there are no ground truth data to be validated against for most of the trip-related metrics.** Therefore, frequently household travel surveys are used for validation purposes. Although this represents a very important first step in the right direction, it is worth noting that the inferred results can have a great number of errors at the individual level, even though a high level of accuracy may be observed at the aggregate level. A number of approaches may be utilized to counter this lack of ground truth, including, for example, the use of simulation data (Chen et al., 2014), collection of small sample GPS/survey data, and using experiments and models to understand the effects of data properties (e.g., locational accuracy and temporal sparsity) on the metrics of interest (e.g., trip rate). Further investigations are critically needed to validate the results generated by big data sources.

6. How do we make useful data via big and small data fusion?

It is clear that there are advantages and disadvantages of big and small data, as well as different types of big and small data. In fact a unique aspect of big data is their continuous and dynamic nature, meaning that they are potentially available at any time and at any place. This is in stark contrast to the small travel survey data that are static, capturing travel patterns on a

typical day once every 5 to 10 years². The static nature and small sample size of travel survey data limit their usefulness for long-term (usually 20- to 30 years) demand forecasts, as well as for assessing many short-term and equally important policy and operations scenarios that arise from time to time.

As an example, understanding anonymous travel patterns in corridor management is critical not only for operations purposes (e.g., evaluating the effectiveness of tolling and other control strategies such as ramp metering) but also for policy evaluation and adjustment (e.g., understanding how different users and communities are affected by the control strategies provides a basis for policy evaluation and adjustment). Big data, because of their dynamic and continuous nature, can be leveraged to provide answers to these important questions. This is the case especially when the big data are integrated with other data, including, for example, household travel survey data, census data, flow data (e.g., travel volumes and speeds from loop detectors), and license plate data that are already collected by state or local departments of transportation (DOTs). This **data fusion** exercise will not only result in useful data that leverage the advantages of diverse data sets, but will also move us toward more real-time, continuous management of our transportation facilities on the basis of the principles of efficiency, equity, and safety. The realization of this vision requires the development of sound data fusion frameworks and methodologies and their validation (beyond a simple combination of the datasets from different data sources), which are currently lacking.

1.2 Recommendations

The ubiquity of passively generated data promises to transform the landscape of transportation planning, from understanding travel patterns to transportation model development and policy evaluation. There has been an explosion of studies using big data to tackle problems in transportation planning. Transportation agencies across the country increasingly find themselves having to make various decisions regarding the purchase and use of big data and their derived products. However, there is a dearth of information about the data themselves (such as data accuracy and representativeness). While every case is different and likely requires a unique evaluation on its own, we offer some general short-term and long-term recommendations based on the analysis conducted in this research.

Ask Questions

This study showed that it is critical to understand the data and their properties, as they directly affect variables of our interest, such as trip rates and OD patterns and the interpretation of analysis results. It is important to ask data providers questions about how the data were generated, what positioning technologies (or a combination of them) were used to locate the devices, what events triggered the recording of the data, and whether there are any reports available on the properties of the data (e.g., locational accuracy and temporal sparsity, among the zeroth-order metrics proposed in this study).

² Most travel surveys are conducted once every 10 years.

Conduct Pilot Tests

Ask for a sample data set from the provider so that analyses can be done on the sample data to further understanding of the data and their derived metrics. The three-order analysis framework proposed in this study can be used to calculate various characteristics. Results from the pilot tests can be compared with those from other studies (such as those from this study) for better understanding of consistency and stability.

Create Benchmark Data Sets and Test Results on the basis of a Common Framework

The use of big data for transportation planning purposes is at its infancy stage. Data representativeness is of critical importance for various types of transportation studies. Therefore, a broad, systematic understanding of such data is urgently needed for big data to reach their full promise in transportation planning.

Develop a common framework

One way to achieve this is to establish a central inventory database in which various benchmark data sets can be created on which metrics can be calculated on the basis of a common framework. This will allow comparisons across different data sets in different geographies, enhancing our understanding of different applications.

Reconciliation of various data sources

While this report does not touch upon other important datasets that are often used in transportation planning applications (e.g., data on traffic flows and transit ridership data), it is important to recognize each dataset (big or small, conventional or emerging) captures a particular view of a transportation phenomenon at a particular scale (both temporally and spatially). In other words, not a single data set will have all the advantages that trump all other datasets, big or small. As an example, it is clear that big data, like the ones studied in this research (from mobile phones and apps), lack the rich behavioral and socio-demographic information that traditional small data sets (e.g., household travel survey data and census data) have. Without this information, it is impossible to answer critical questions related to geographical or socio-demographic equity. Therefore, any decision on which datasets to be used and how they may be reconciled together hinges upon knowing what particular transportation phenomenon to be captured and what datasets will help capture aspects of the phenomenon of interest. In some cases, rigorous data fusion techniques will need to be developed in order to integrate various data sets together by leveraging their unique advantages. In other cases, individual datasets can be used to capture different aspects, which together form a complete story explaining a transportation phenomenon of interest.

Investigate and understand the evolutionary nature of big data and the impact of changes on the use of big data

It is of paramount importance that we recognize the evolving nature of big data. As demonstrated in our study, as technologies evolve, the nature of big data (as measured by

properties such as the zeroth order ones) changes, too. And this can directly affect the estimation of trip-related characteristics. Technologies will continue to evolve. The advent of autonomous and connected vehicles, for instance, will provide a whole suite of new data related to the car, its driver and passengers, surrounding traffic, and the immediate environment. The new data will not only help us gain new insights into transportation planning, operations, and safety analyses, but will also raise new questions about the data themselves, their properties, and how they may affect the derived trip-related characteristics, analysis results ensuring representativeness, equity and fairness, and impacts on our policies

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.

2.2 Acknowledgments

The FHWA would like to acknowledge the assistance of two metropolitan planning organizations (MPOs) that 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. Chenxi Liu from the University of Washington also contributed to early versions of Chapter 4 of this report.

2.3 From Travel Survey Data to Passively Generated Emerging Data

Since the 1950s, household travel surveys, as an important source for transportation planning applications, have gone through significant changes in survey instrumentation, methods of survey administration and sampling, and consequently response rates and sample sizes (Stopher, 1996). The earliest travel surveys were conducted via personal interviews, with a response rate of between 80 and 90 percent (Stopher, 1996), followed by mail-out and mail-back surveys and telephone interviews via random digit dialing. The last two decades have witnessed a rise of web-based surveys or a combination of paper-, web-, and phone-based surveys. In recent years, the prevalence of smart mobile devices has prompted the development of smart phone-based travel survey applications (Cottrill et al., 2013; Fan et al., 2013; Liao et al., 2017). The response rates for travel surveys have dropped to about 25 percent in the last decade (Stopher and Greaves, 2007), with sampling rates ranging from less than 1 percent for large urbanized areas to less than 3 percent for small ones (Stopher and Greaves, 2007).

Since travel surveys are actively solicited and rely entirely on self-reporting by respondents, it is widely recognized that short trips, trips made by non-motorized modes, and/or first- and last-mile trips are often ignored (Wang et al., 2019). There is also increasing nonresponse, either because targeted households do not respond to an entire survey or to specific items in a survey. Related to the nonresponse issue is the non-representative concern (Wang et al., 2019). The fact that nearly all surveys capture only a tiny fraction of the population also adds to the non-representative concern.

The above concerns have greatly motivated interest in using passively solicited data to supplement or even replace household travel surveys (Chen et al., 2010). As defined by Chen et al (2016), passively solicited data are those generated by non-transportation-application related primary purposes (e.g., billing, app use) but that can be potentially used for transportation planning. Examples include mobile phone data, vehicle GPS data, app-based data, social media data, etc. All such data include spatial and temporal information, which forms the basis for identifying people's mobility patterns; they differ significantly from travel surveys (or

actively solicited data) in a number of aspects.³ Therefore, it is naturally expected that the resulting data will have unique characteristics that distinguish them from survey data. In fact, because different kinds of passively solicited data (e.g., mobile phone, vehicle GPS, app-based data) are generated through different processes, all likely possess their own characteristics.

This report continues from a previous report by the authors (Chen et al., 2017) in which characteristics associated with mobile phone data and vehicle GPS data were investigated and compared with household travel survey counterparts. More specifically, we investigated the characteristics of an emerging passively solicited data set: app-based data. Unlike mobile phone or vehicle GPS data generated through a single-sourced positioning technology (cellular triangulation for the former and GPS for the latter), app-based data are multi-sourced, meaning that a combination of technologies (GPS, WiFi, cellular triangulation, Bluetooth, etc) are used to position the devices. This suggests that different methodologies must be used to extract travel patterns. The study also showed that the app-based data differed significantly in a number of ways from the mobile phone data and vehicle GPS data, as well as data from household travel surveys.

The rest of this report is organized as follows. In Section 3, analysis of emerging data source collected via mobile apps (app-based data) is discussed. The app-based data were compared with travel survey data that were collected in the same study region and time. Discussions of the characteristics of the app-based data, as well as recommendations on their use, are provided. Section 4 provides a summary of other data sources from emerging technologies and systems in transportation and their potential applications. These include data from connected and automated vehicles (CAVs) and new shared mobility services. With an understanding of these emerging data sources, Section 5 provides a discussion of the development of a data fusion framework, with a goal of producing better quality data and/or more complete data for given transportation planning or operational applications.

³ Since passively solicited data are not generated through probabilistic sampling plans, terms such as response rate and sample size are no longer meaningful.

3.0 App-Based Data

This section discusses the analysis of an emerging data source collected via mobile apps. The analysis followed the framework developed by the project team in the first phase of the project (Chen et al., 2017) to analyze data properties of different orders. (“Order” here refers to the number of activity locations, so that zeroth order properties refer to those related the data themselves with no activity location derived, and first and second orders refer to single- and two-activity location(s) (or trips), respectively.) More details of such properties for the zeroth, first, and second orders can be found in Table 1.⁴

Table 1. An overview of the analysis framework.

Order	Properties	Contents	Variations
Zeroth Order	General description	Time period/counts of observations/counts of unique IDs/spatial distribution of data	N/A
Zeroth Order	Location accuracy	Statistics of location accuracy of all data	Distribution and cumulative distribution
Zeroth Order	Temporal distribution of observations	Distribution of observations across one day/number of observations/ number of observations per ID	Daily and weekly
Zeroth Order	Inter-day temporal sparsity	Distribution of the number of days observed/distribution of life span of unique IDs	N/A
Zeroth Order	Intra-day temporal sparsity	The temporal resolution of a trajectory/number of trajectories revealing their locations across a day	Daily and weekly
First Order	Comparison between inferred home census tracts and census data	Spatial comparison and correlation analysis between inferred home census tracts and census data	N/A
First Order	Activity duration	Distribution of observed times at inferred activity locations	N/A
First Order	Spatial distribution of trip ends	Spatial distribution of extracted trip ends	Weekdays vs weekends
Second Order	Trip rates	Distribution of trip rates per day per anonymous user	Daily variations
Second Order	Departure/arrival times	Distribution of departure/arrival times of trips	Time of the day; weekdays vs weekends
Second Order	Trip length/Travel times	Distribution of trip length/travel times	Cumulative distribution; weekdays vs weekends
Second Order	Origin-destination demand	Correlations between estimated OD demands and PSRC OD demands; spatial distribution of OD demands	N/A

⁴ Note that the list of properties is not identical to that in the first phase of the project (Chen et al., 2017), as some properties were no longer applicable to the data sources analyzed in this project.

3.1 General Description (Zeroth Order)

The app-based data analyzed in this study was provided by Cuebiq, location intelligence and measurement company that supplies software development kits (SDKs) to mobile app developers, providing a privacy-compliant path for anonymous users to opt in to share location data. The kits allow developers’ apps to use “Location-Based Services.” About 180 apps were included, with functions such as shopping, travel, and navigation. The data provider was said to collect app-based data from 61 million monthly active anonymous smartphone users (about 20 percent of the U.S. population) who opted in to share location data. Each observation in the data set contained the device ID of an encrypted anonymous mobile device, a time stamp, a location record (in the form of a pair of latitude and longitude coordinates), and the associated location accuracy in meters. Table 2 gives a synthetic sample of the observations.

The time interval between two consecutive observations varied, depending on the usage pattern of the mobile phone applications that contributed to data generation, as well as the data collection frequency set by the data provider. The location accuracy of observations varied as well, depending on the positioning technology used when the data were collected. This could range from a few meters when the GPS chip was on to more than 20 meters when apps recorded locations using Wi-Fi proximity, assisted GPS (AGPS), Bluetooth proximity, etc (Schewel, 2017). In the extreme case scenario, the accuracy could be hundreds or even thousands of meters off when observations were recorded using cellular towers.

Table 2. A synthetic sample of app-based data

Time stamp	Device ID	Latitude	Longitude	Location accuracy (meters)
1491398264	4ab844ff98c206b8d7	47.9205809	-122.2535626	5
1491403834	4ab844ff98c206b8d7	47.9229781	-122.2903396	25
1491403961	4ab844ff98c206b8d7	47.9222743	-122.2998663	60
1491412669	4ab844ff98c206b8d7	47.8994576	-122.2915348	60
1491412963	4ab844ff98c206b8d7	47.8856073	-122.2908753	300
1491413263	4ab844ff98c206b8d7	47.8850917	-122.2806468	1399

The app-based data used in this report were collected between April 4, 2017, and June 5, 2017 (63 days). These data were spread out within the Puget Sound region (PSR) among four counties: King, Kitsap, Pierce, and Snohomish. Figure 1 shows the spatial distribution of observations on a sample day (May 9, 2017). It can be observed that the majority of the observations were on the west side of this region (the Seattle area) because of its higher population density.

Figure 2 provides a zoom-in view of the observations in the central PSR on a typical weekday during the evening peak (17:00 PM to 18:00 PM), showing clusters of observations along major

highways. This suggests that observations during peak hours were closely linked to travel activities because of the way the data were collected; according to the data provider, anonymous users' locations were sampled more frequently when they were moving than when they were static.

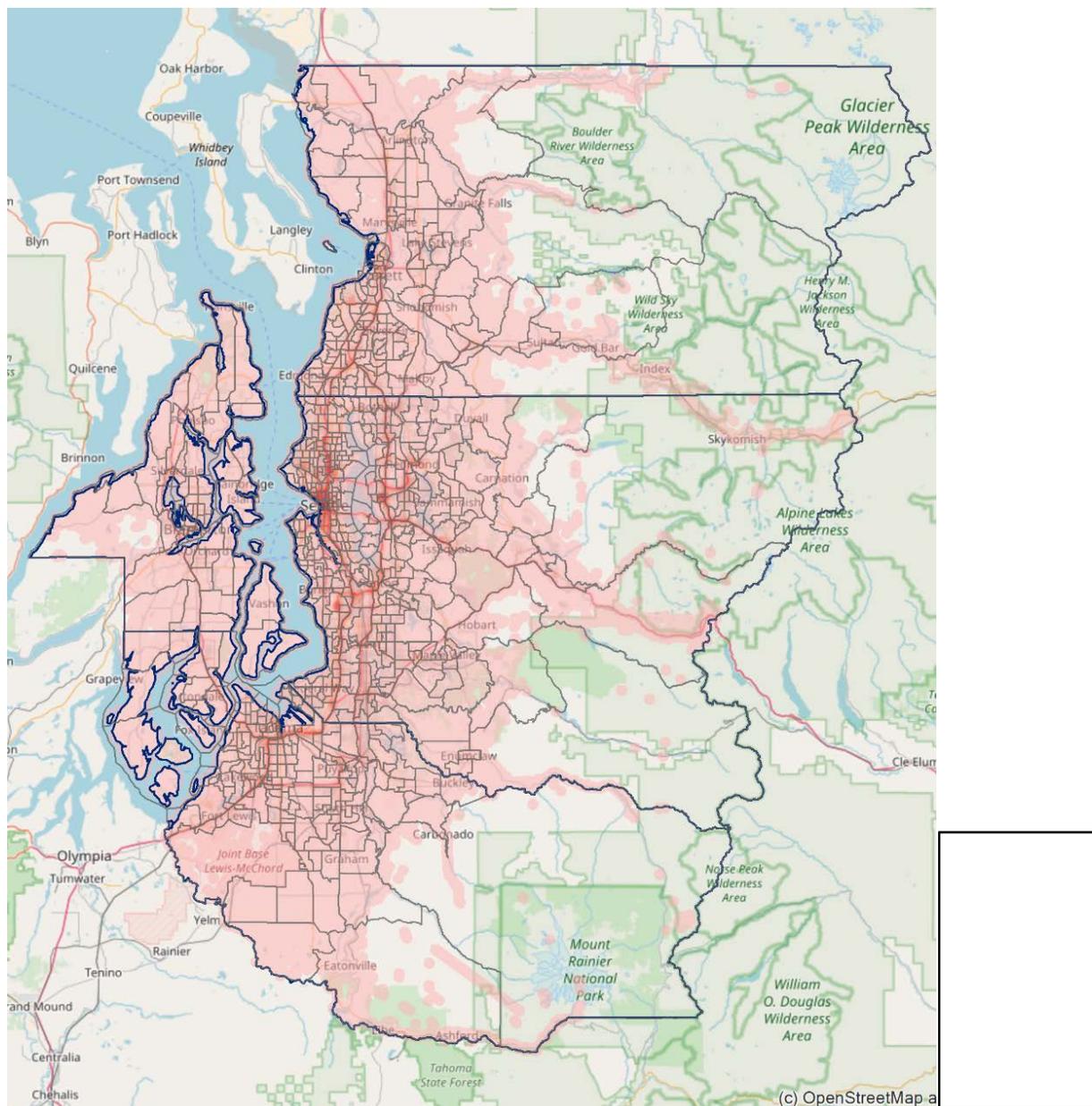


Figure 1. Map. Spatial distribution of observations

Source: World Topographic Map

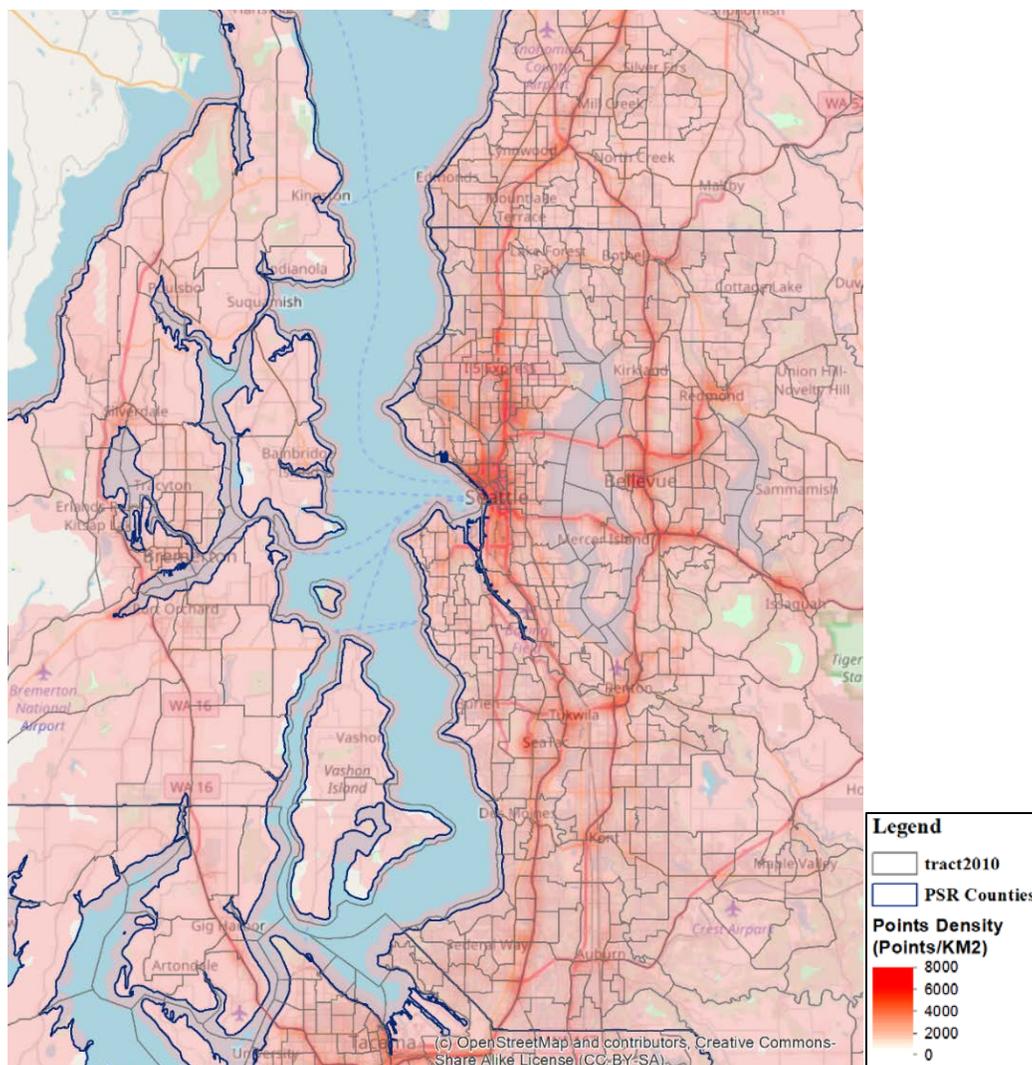


Figure 2. Map. A zoom-in view of the central PSR during the evening peak of a typical weekday

Source: World Topographic Map

Spanning 63 days, the data set contained 462,401 unique device IDs producing 563,038,663 observations. This 462,401 number was equivalent to about 12.8 percent of the population in the Puget Sound region (3,798,902 persons), if each device were considered a resident. Note, however, that the underlying population of this data set would also contain many non-residents, such as those who visited or passed by the region, as well as those who carried multiple devices.

Figure 3 shows the distribution of the time intervals between two consecutive observations (of one ID within a day). Observe that the majority of the observations (84 percent) had time intervals of less than 60 seconds, sharing some similarity to vehicular GPS data (Chen et al., 2017).

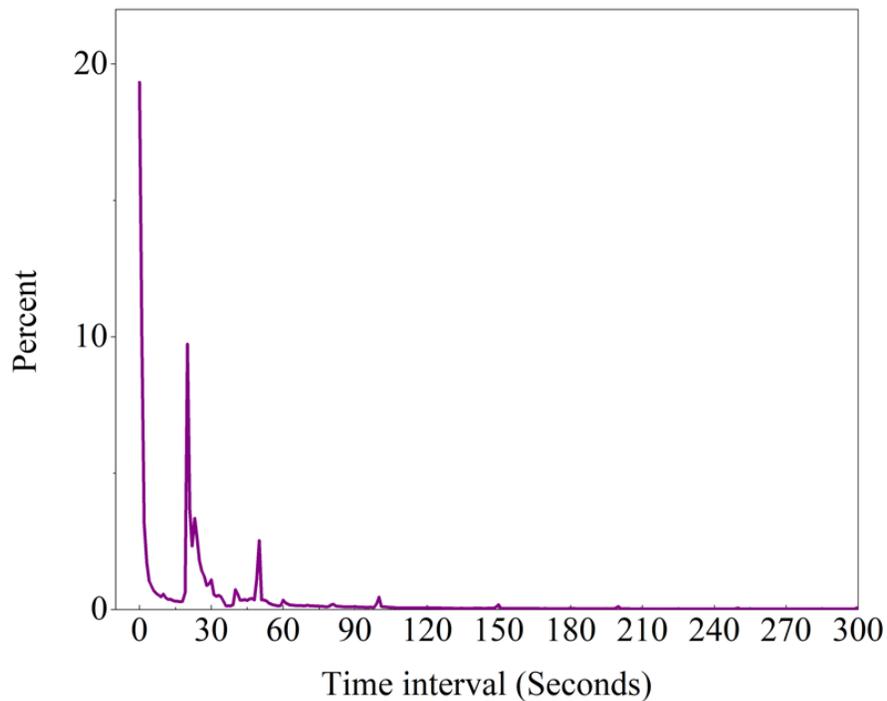


Figure 3. Graph. Time interval distribution between two consecutive observations

3.1.1 Location Accuracy

As mentioned earlier, observations were generated by using various positioning technologies. Figure 4 shows the distribution of locational accuracies in the data set and Figure 5 is the cumulative distribution. One quarter of the observations had an accuracy of less than 5 meters; about 85 percent of the observations had an accuracy of less than 100 meters; 93 percent of the observations had an accuracy of less than 1,000 meters; and a small percentage of observations (about 2 percent) had an accuracy level exceeding 2,000 meters.

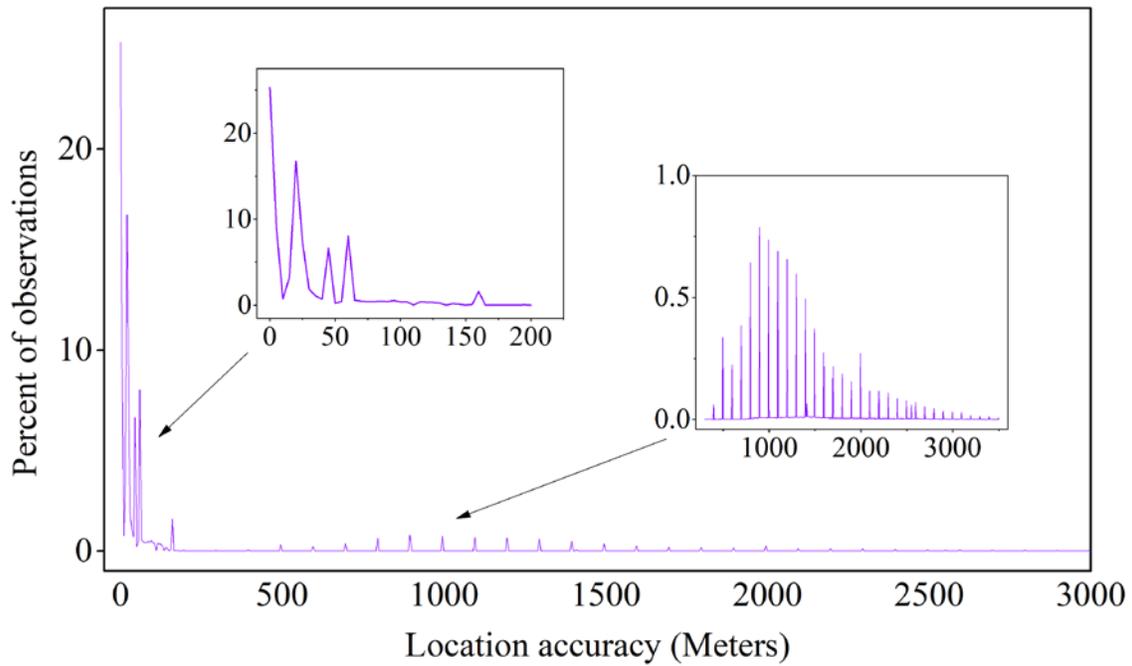


Figure 4. Graph. Distribution of location accuracy

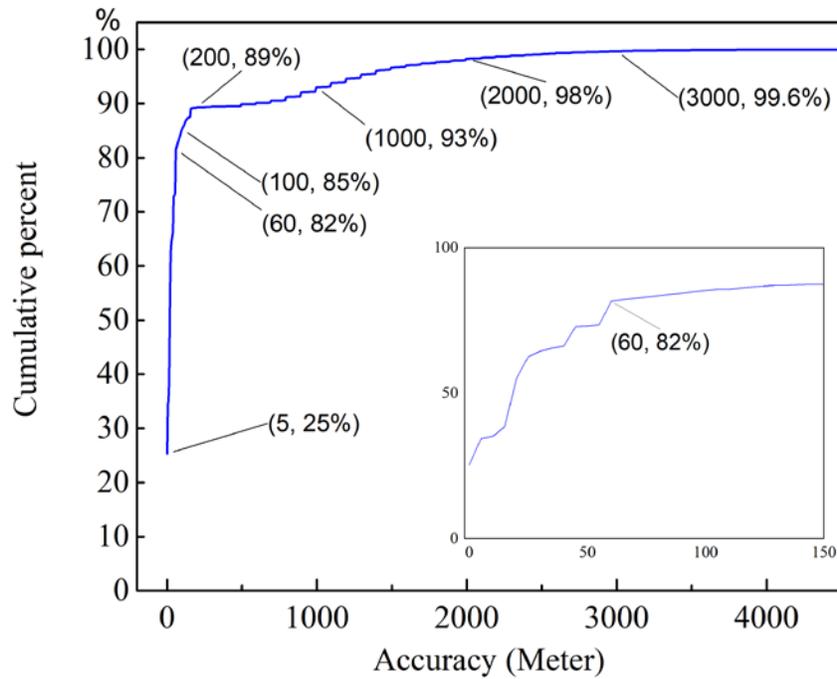


Figure 5. Graph. Cumulative distribution of location accuracy

3.1.2 Temporal Distribution of Observations

Temporal Daily Distribution of Observations

Figure 6 shows how observations were distributed over a day, with comparisons across different days of a week. Two peaks were observed (morning peak and evening peak) on weekdays, and one mid-day peak was seen on weekends. In addition, the morning peaks on weekdays were lower than the evening peaks, suggesting a lower level of apps usage in the morning. The distribution of observations for Fridays was higher after 10:00 AM than on weekdays.

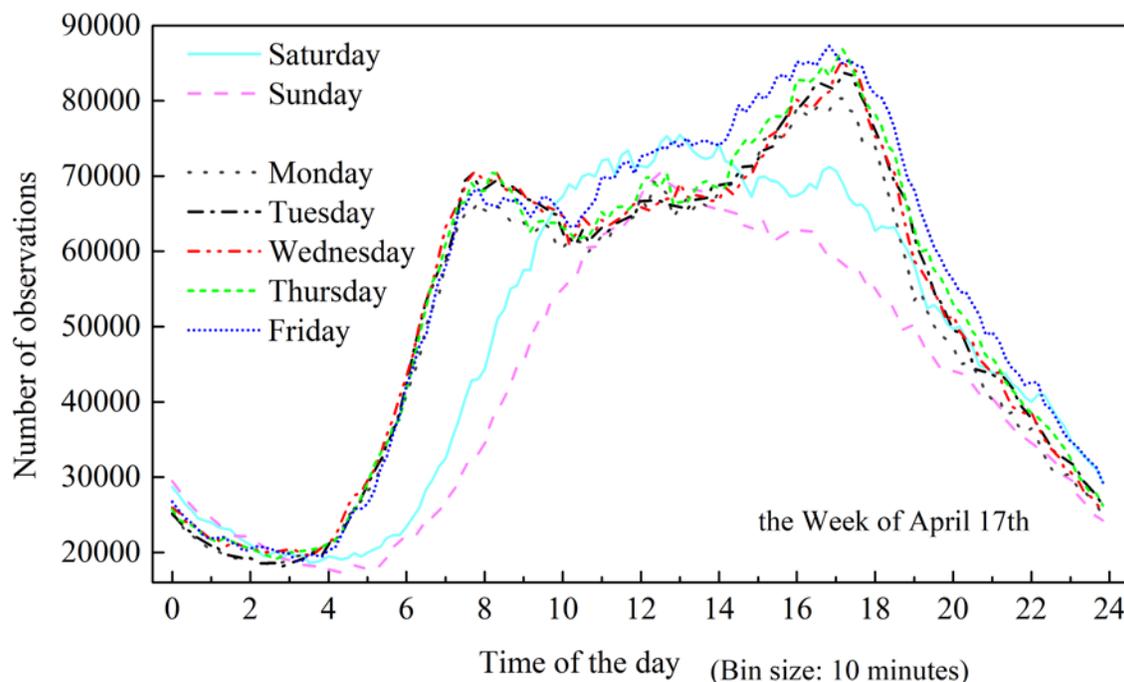


Figure 6. Graph. Distribution of observations within a day (the Week of April 17th)

Temporal Weekly Distribution of Observations

On average, there were more observations on weekdays than on weekends (Figure 7). Fridays had the most number of observations, while Sundays had the lowest, except for Memorial Day on May 29th.

Figure 7 also shows that the number of daily observations suddenly increased on May 9, 2017. This may be attributed to the increase in the number of mobile apps from which data were collected. More discussion on this can be found in Section 3.4, May 9th Data Shift.

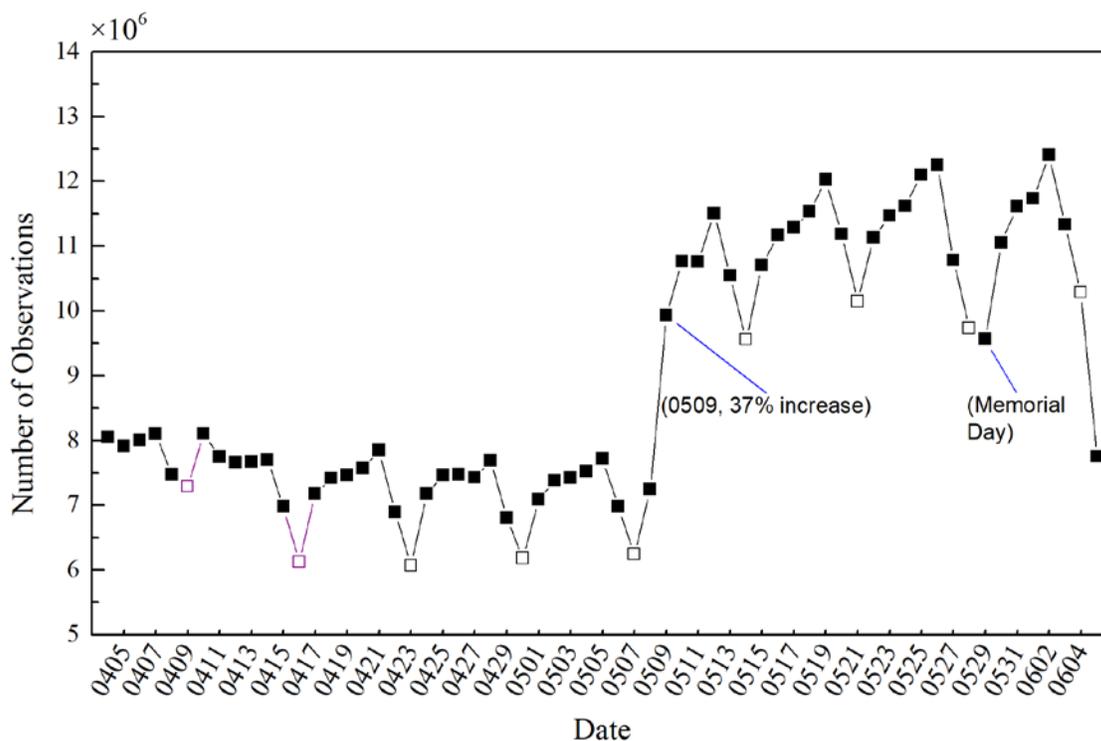


Figure 7. Graph. Weekly pattern of observations (Sundays are in open box)

Location Frequency Update

As shown in Figure 3, the time interval between two consecutive observations could be several seconds, hours, or even days, suggesting that the update frequencies of anonymous users' locations were not uniform. How regularly anonymous users' locations were updated within a day was then investigated. First, a day was evenly divided into multiple time intervals, each 30 minutes. For each time interval, we then checked to see whether an ID/trajectory revealed its location at least once. Figure 8 shows the distribution of the presence of IDs/trajectories at different times of a day. These were called location updates.

Weekdays shared similar patterns of having three peaks during a day: morning, noon, and evening. This suggests that on weekdays, the locations of the IDs were more likely to be revealed during these peak hours than during other times. More-IDs were also observed on Fridays than on other weekdays. On the other hand, weekend days followed a unimodal distribution, peaking in the early afternoon.

The location update patterns in Figure 8 are similar to the temporal distribution of observations shown in Figure 6. However, the comparison between the two figures does show one clear difference: the morning and evening peaks in Figure 8, with location updates, are less striking than those in Figure 6, with only observations, especially for the morning peak. This suggests that some anonymous users' locations were updated at a high frequency during the morning

and evening peak periods, leading to the magnified peaks in Figure 6. As shown later in Section 3.3.2, this phenomenon also affects the departure time distribution within a day.

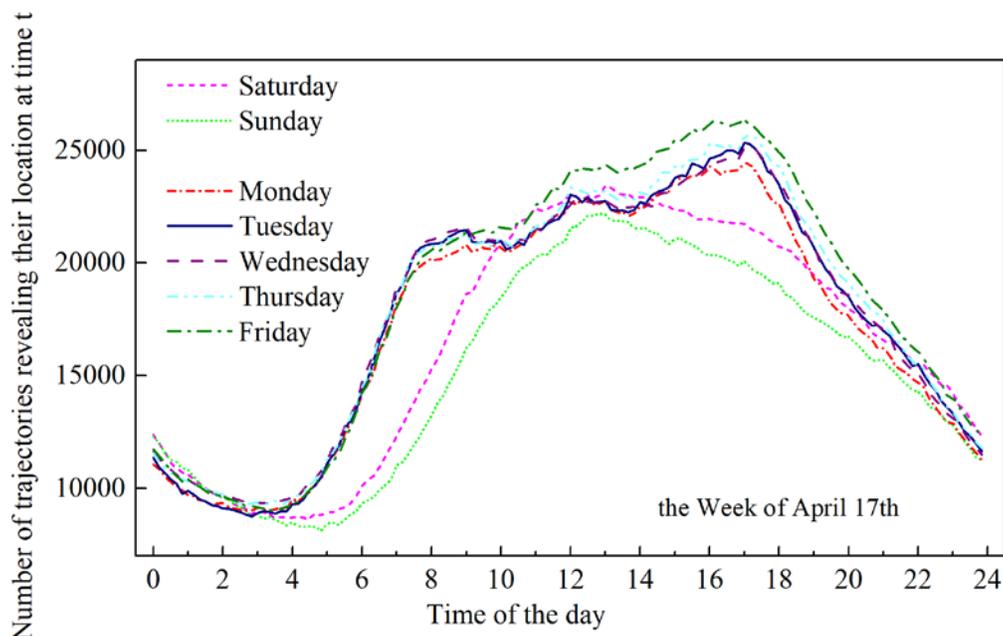


Figure 8. Graph. Percentage of trajectories with their locations revealed at a time of the day, comparing different days in a week.

The evolving pattern of a holiday was also investigated (Memorial Day, May 29, 2017), as shown in Figure 9. The holiday temporal pattern was clearly different from the weekday one, showing a unimodal distribution that peaked around noon.

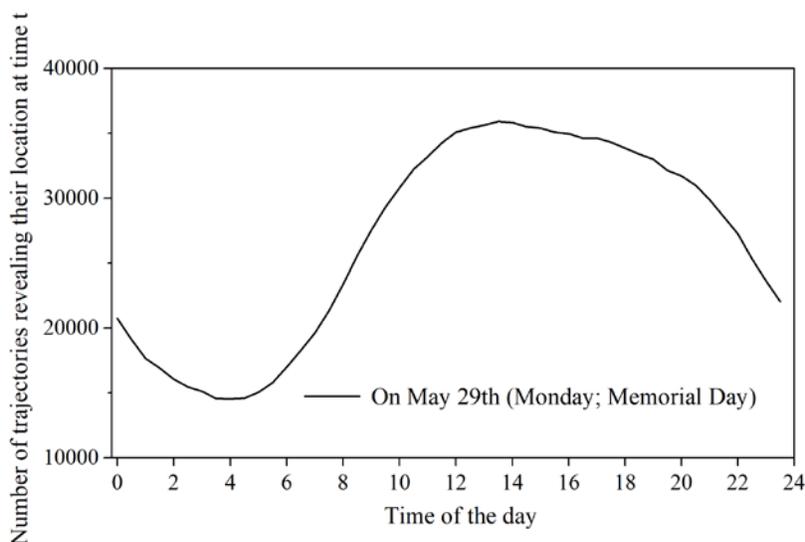


Figure 9. Graph. Percentage of trajectories with their locations revealed at different times of a holiday.

3.1.3 Temporal Sparsity

In comparison to actively solicited data (e.g., household survey data), passively generated data such as app-based data rely on anonymous users' activity patterns and the update frequency of the data provider (that collects the data), and therefore they can be temporally sparse. Specifically, for the app-based data, observations are collected only when the apps are running. In this report, temporal sparsity was investigated via two measures: inter-day and intra-day sparsity, quantifying how an anonymous user's observations were distributed across different days and across different times within a day.

Inter-Day Temporal Sparsity

Two factors potentially contributed to inter-day sparsity: 1) visitors or passersby may have appeared in the data set for only a short period of time (one day or a few days); and 2) residents in the region may not have continuously used the included apps or may have travel out of the region during the study period. Figure 10 shows the distribution of the *life span* of each anonymous user (ID), which was defined as the difference between the first and last day that an ID was observed. For example, an ID that had its first observation on May 1st and its last observation on May 3rd had a life span of three days. It can be observed that 33 percent of IDs had a life span of only one day, and 53 percent of IDs had a life span of less than one week. This implies that a significant fraction of anonymous users in the data set were either not residents or were residents who did not use their apps frequently. On the other hand, 10 percent of IDs had a life span for the entire study period (63 days).

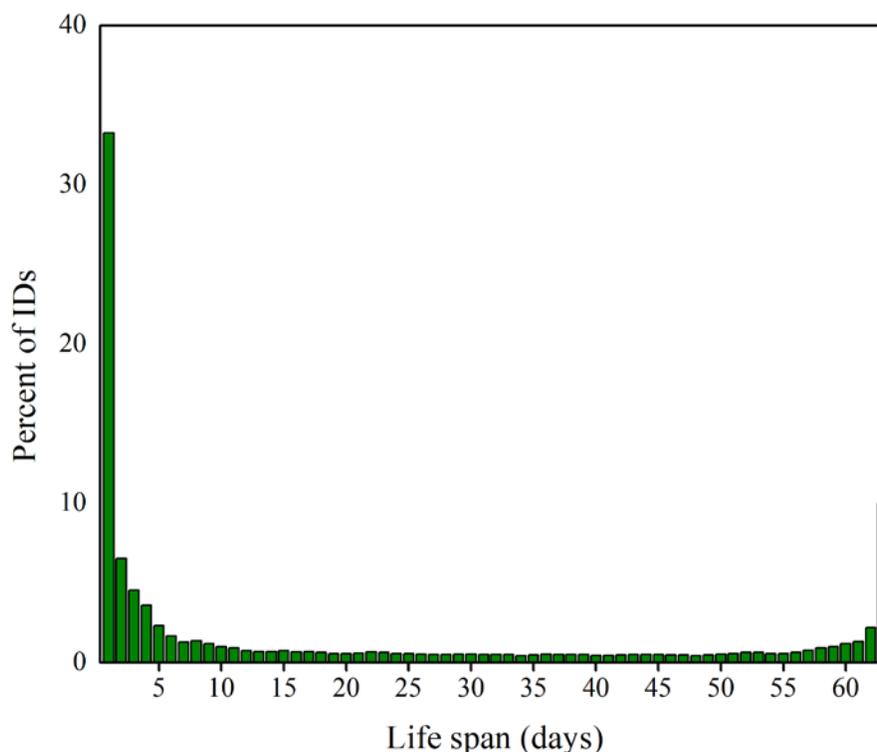


Figure 10. Graph. Inter-day sparsity (distribution of life span of unique IDs)

Figure 11 shows the distribution of the number of days observed for each ID. As seen in the figure, most of the IDs were observed for only a few days: 66 percent of IDs were observed fewer than seven days. Differences between the distribution of life spans and the number of observed days can also be observed, suggesting that most anonymous users did not use their apps continuously. For example, although 12 percent of IDs had a full life span (63 days), only 3 percent of IDs had observations every day⁵. Additionally, while 12 percent of anonymous users were observed for two days, only half of them had a two-day life span. This means that the other half of anonymous users had a life span of longer than two days (e.g., an anonymous user may have had her first and last observation on May 1st and 3rd, but no observations on May 2nd, resulting in a three-days life span and two observation days).

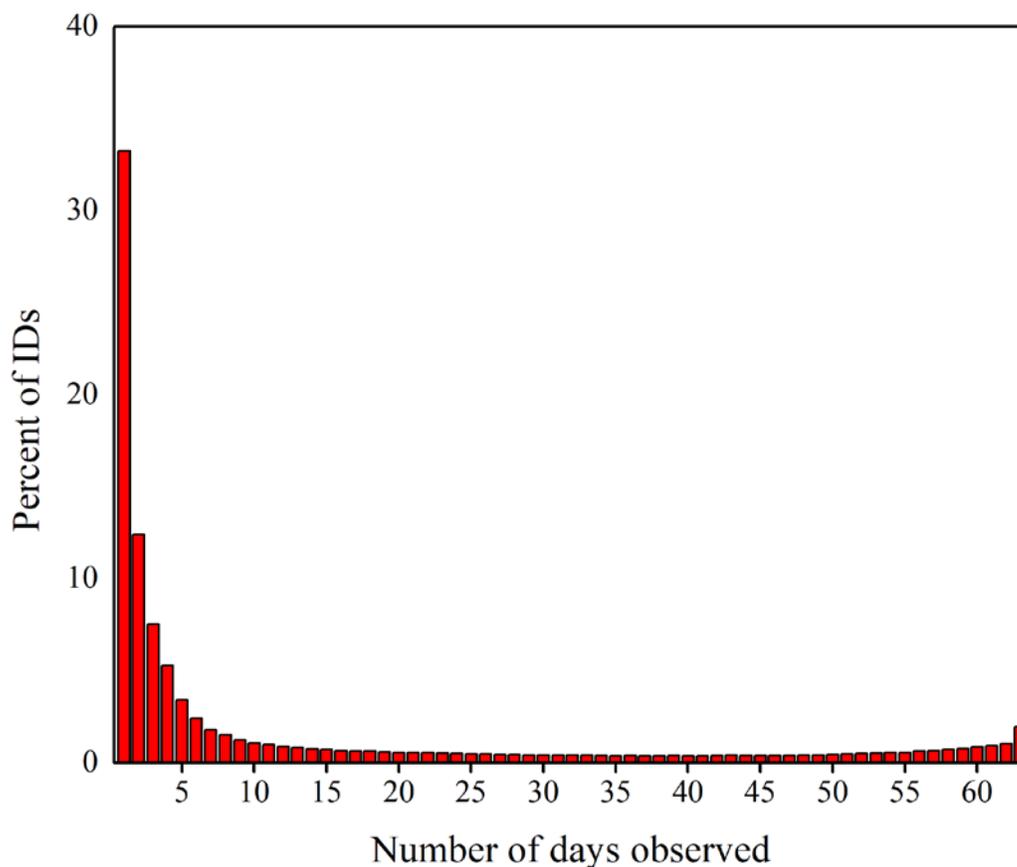


Figure 11. Graph. Inter-day sparsity (distribution of number of days observed)

Intra-day Temporal Sparsity

For the days with data observations, we further investigated how these observations were distributed within each day. Figure 3 shows that most time intervals (84% percent) between two

⁵ A spatial distribution of these users is provided in Appendix C.

consecutive observations were less than 1 minute. However, this does not necessarily mean that anonymous users' locations were being recorded continuously throughout the day. Instead, as shown in Figure 6, observations were likely to cluster during the morning and evening peak hours while being absent during other times.

To capture the potential data sparsity within the day, a day was divided into 48 time-slots (30 minutes for each slot). For trajectory_{id} (the sequence of observations of user *i* on day *d*), we defined its temporal resolution ϕ_{id} as the number of time slots in which this user was observed at least once (i.e., $\phi_{id} \in [1, 48]$). Note that only the days with observations were considered. Figure 12 shows the distribution of ϕ_{id} , for all trajectories. Here, one user may have contributed more than one trajectory, depending on the number of days that he/she was observed. The median was 10 slots, indicating that half of the trajectories had no more than 5 hours, with their locations revealed on average within a day. Only 0.4 percent of trajectories had their locations observed at each time slot (i.e., every 30 minutes).

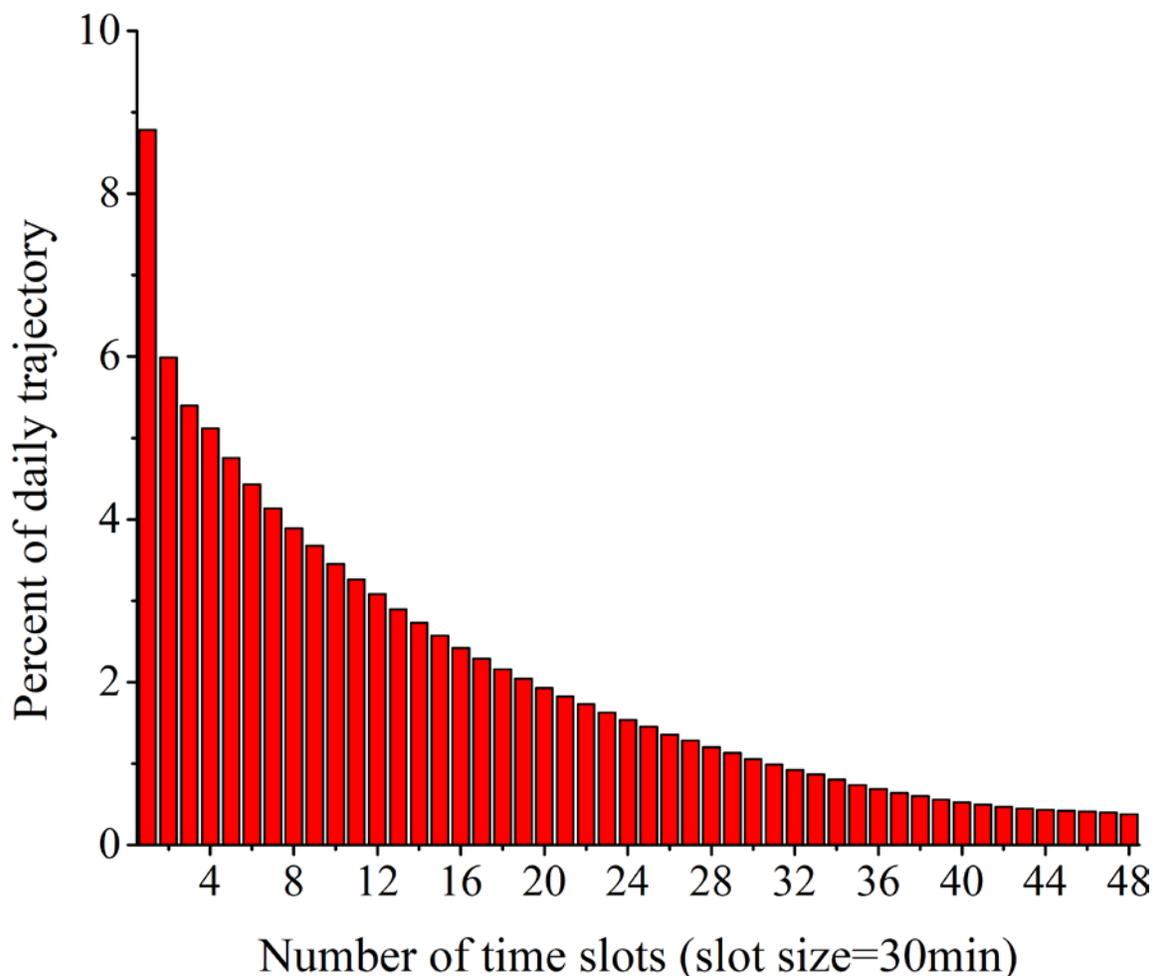


Figure 12. Graph. Distribution of temporal resolution of all (daily) trajectories.

3.2 First-Order Properties

In this section, the properties of stays identified from the data are further explored, including the time spent at stays (activity durations), their spatial distribution, and associated departure times. Passively generated for non-transportation purposes, app-based data need to be processed to identify stays⁶. In this study, the properties of extracted stays and trips were compared with those from a household travel survey data. The survey data were collected by the Puget Sound Regional Council (PSRC) *during the same period* that the app-based data were generated in the spring of 2017. The survey data contained 6,254 persons residing in 3,285 sampled households in the Puget Sound region (Michalowski, 2017).

3.2.1 Identifying Home Census Tracts

Since the PSRC household travel survey was a sample of residents only, the non-residents in the app-data needed to be removed first for comparison. The home census tract for an anonymous user was defined as the census tract with the most frequent visits during the evening (22:00 PM to 6:00 AM the next day). In this study, this was defined as at least eight visits during the entire two-month study period, representing an average of at least one visit per week. Figure 13 shows that the spatial distribution of inferred home census tracts was similar to the population estimated by the American Community Survey (2015 American Community Survey). With a Pearson correlation coefficient of 0.91 at the census tract level, Figure 14 shows that the estimated density of home census tracts scaled well with the population represented by the census.

⁶ Details on how to identify stays and extract trips from the app-based data are provided in Appendix A.1

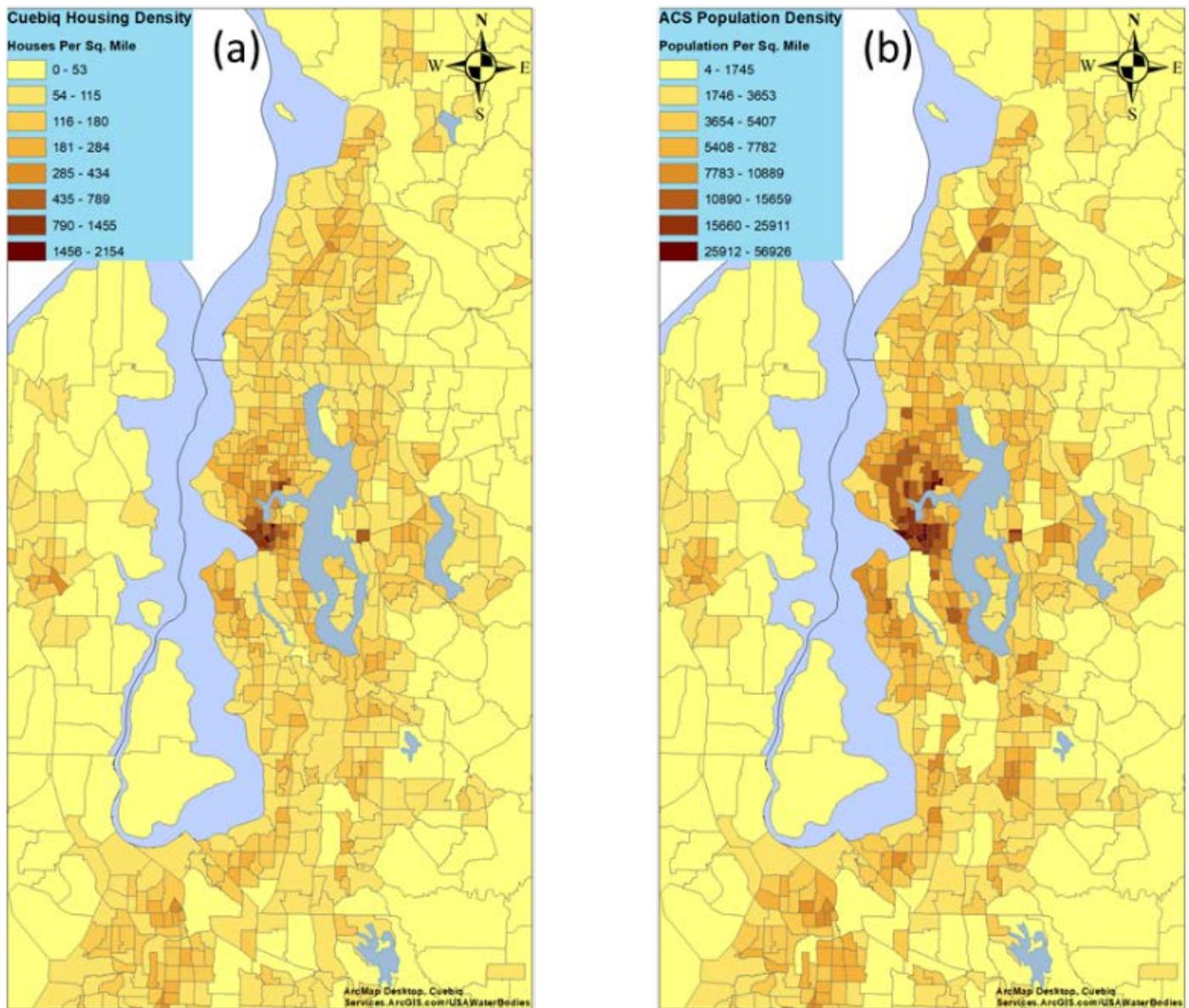


Figure 13. Map. Comparison between home census tracts inferred from the app-based data and the population from the census. (a) Inferred home density at census tract level and (b) Census population density.

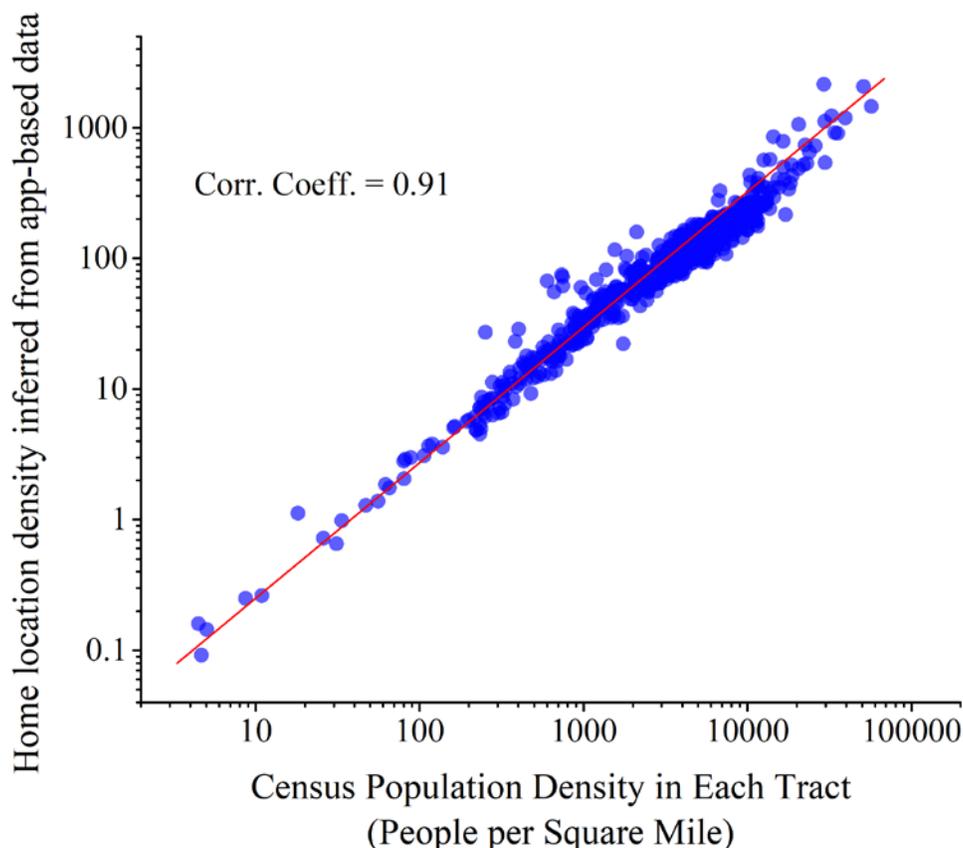


Figure 14. Graph. Correlation between inferred home census tracts and census population at census tract level.

Among all the anonymous users in the app-based data, about 24 percent of them (110,889) had a home census tract inferred, equivalent to about 3 percent of the population in the Puget Sound region. This is much higher than the 0.21 percent sampling rate for the PSRC household travel survey. The sample sizes can be further compared at the census tract level. For each tract, the sample size of the app-based data was calculated as the ratio between the number of inferred residents and the population residing in that tract. Figure 15 compares the distribution of tract-level sampling rates for the app-based data with the distribution from the survey data. It shows that for most tracts, the sampling rates for the app-based data were larger than those of the survey data. More specifically, 84 percent of the sampling rates for the app-based data ranged between 1 percent and 5 percent. On the other hand, about 89 percent of the sampling rates for the survey data were less than 1 percent. Interestingly, the sampling rates for the app-based data showed an extended triangle distribution.

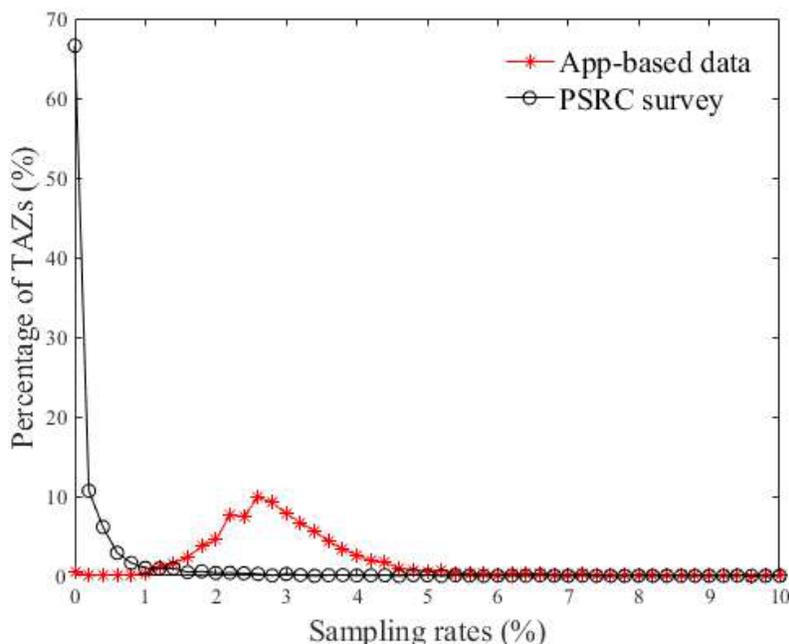


Figure 15. Graph. Distribution of scaling factor

3.2.2 Activity Duration

Users' trajectories are typically filled with travel activities on the road and stays at activity locations. Activity duration describes how long a subject stayed at certain place for an activity. Because of the temporal sparsity discussed above, as demonstrated in Figure 16, the observed arrival time \hat{t}_{arr}^i from one activity location i may not be the actual arrival time t_{arr}^i . This is also true for the observed departure time \hat{t}_{dep}^i . Therefore, the observed activity duration, which is defined as the time difference between the observed departure and arrival time, could potentially be an inaccurate estimation of the actual activity duration (often an underestimation, as $(\hat{t}_{dep}^i - \hat{t}_{arr}^i < t_{dep}^i - t_{arr}^i)$). For the same reason, the estimation of travel time is also not accurate (often an overestimation), as is shown in Section 3.3.3.

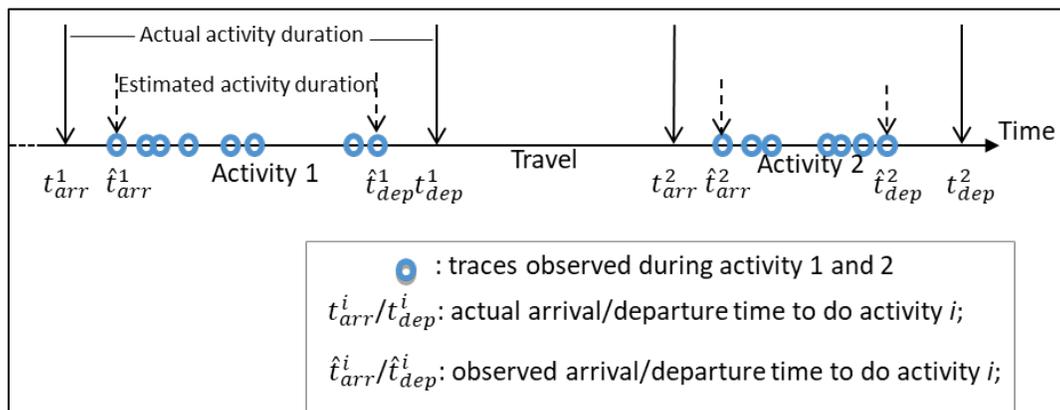


Figure 16. Illustration. Activity duration and a demonstration of the biased estimation.

Figure 17 gives the distribution of activity durations inferred from the data and compared with those from the PSRC survey data. Both distributions showed a consistent decay for activity durations of less than 8 hours. Three observations can be made. First, short stays (of less than an hour) were more represented in the apps data than in the survey data. Second, stays of medium duration (between 3 and 6 hours) were also more pronounced in the apps data than in the survey. Third, stays of long duration (more than 8 hours) were more under-represented in the apps data than in the survey. Several factors may have been at play:

- 1) Short trips (e.g., visiting a coffee shop) during a long stay (e.g., at a workplace) may have been under-reported in the survey data (Wolf et al., 2003).
- 2) Signaling noise in the app-based data may have been falsely identified a single stay as multiple stays with movements between them.
- 3) Underestimated activity duration may have been due to the lack of observations at true activity starting times and ending times (see Figure 16), which was especially the case for the third observation stated above.

The third factor also explains the reason why long trips (in terms of travel times) were over-represented in app-based data (see Section 3.3.4, Figure 25). For the app-based data, no clear differences could be observed in the distribution of activity duration between weekdays and weekends.

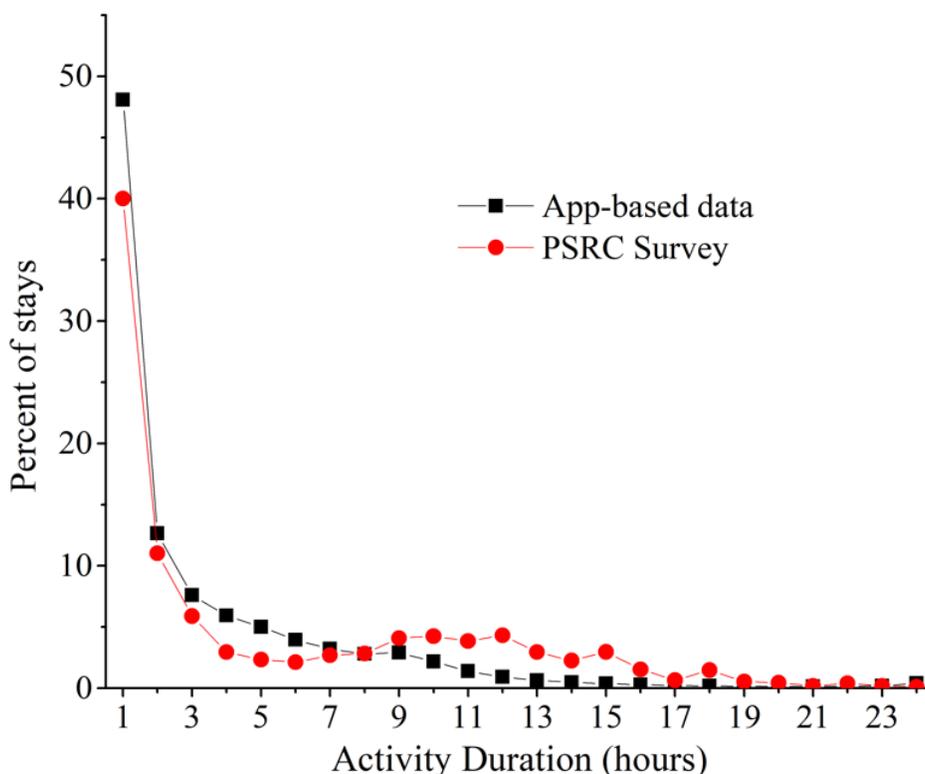


Figure 17. Graph. Activity duration observed from PSRC survey and app-based data

3.2.3 Spatial distribution of extracted trip ends

this subsection shows the spatial distribution of the extracted trip ends. In this report, terms such as trip ends and stays are interchangeable, as we defined a trip as a pair of two consecutive stays. Figure 18 illustrates the spatial distributions of trip ends on a typical weekday morning, showing travel demands concentrated in city centers such as downtown Seattle, Tacoma, and Bellevue.

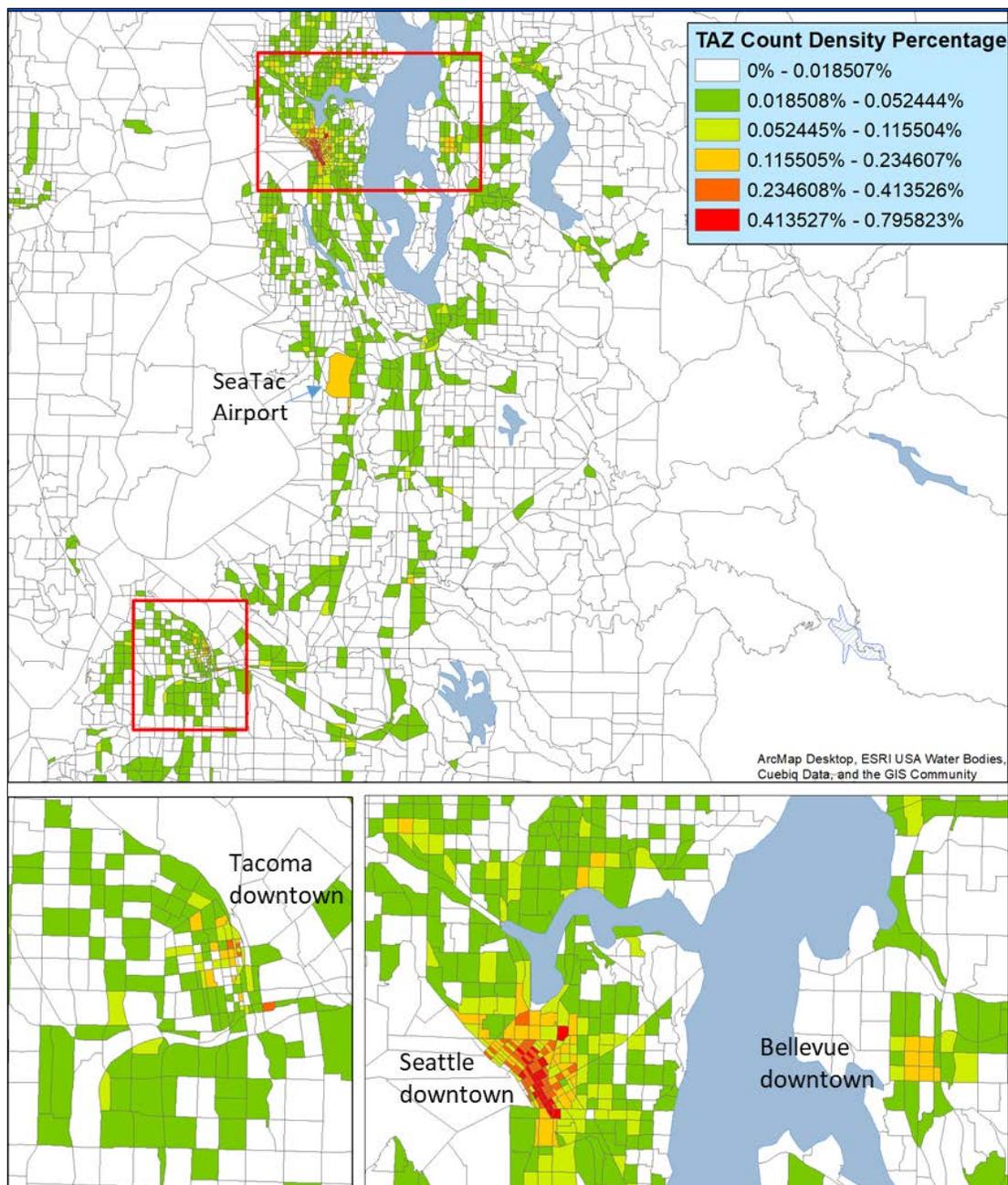


Figure 18. Graph. Spatial distribution of trip ends on a weekday morning.

Figure 19 illustrates the differences in trip ends between weekdays and weekends. More trip ends were observed in downtown Seattle, Bellevue, Everett and Tacoma on weekdays than on weekends (shown in red dots).

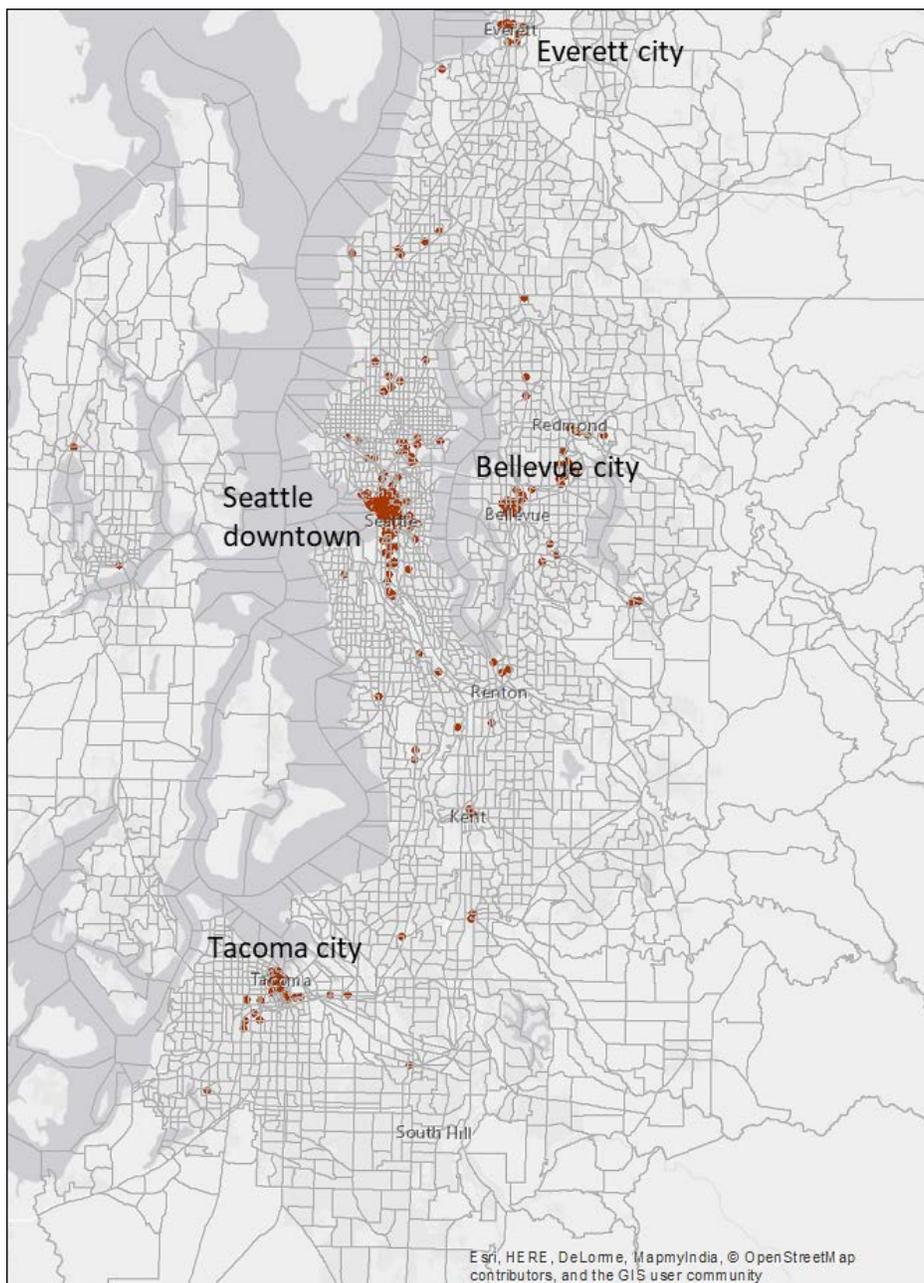


Figure 19. Graph. Spatial distribution illustrating where more trip ends are observed on weekdays than that on weekends (in TAZ)

Source: World Light Grey Base

3.3 Second-Order Properties

Once trip ends had been identified, their connections, or a trip, could be inferred. Furthermore, by deriving individual trips, the origin and destination could be established for each trip. They could then be aggregated to estimate the travel demand for a region.

3.3.1 Distribution of Trip Rates

Trip rate was defined as the number of trips a person conducts in a day. Figure 20 shows the trip rate distribution inferred from the app-based data in comparison to that from the PSRC survey data. It can be observed from the app-based data that 14 percent of users did not generate any trips while approximately 36 percent of users generated one to two trips per day. In comparison, the PSRC survey data showed that: 1) a trip rate of 2 was most frequently observed; and 2) 11.5 percent of user-days had zero trips and about 1.8 percent had one trip.

The mean trip rate was 3.23 for the app-based data, less than the 4.4 estimate from the survey data. Those conducting a single or no trip in a day were overestimated by using the app-based data, while those conducting more than one trip were underestimated. This was likely due to the temporal sparsity issue of the app-based data as discussed in Section 3.1, as some of the trips were not captured in the data (i.e., missing trips). Short trips, whose trip ends were close in space, may also have been missed because of location uncertainty.

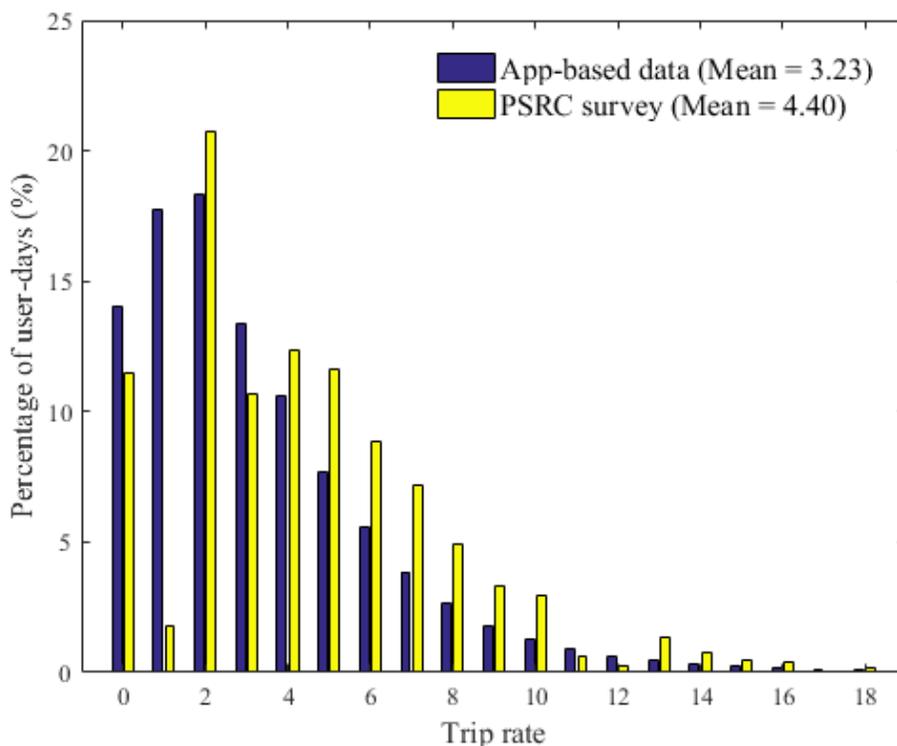


Figure 20. Graph. Distribution of trip rates

The average daily trip rates of the app-based data are shown in Figure 21. Consistent weekly patterns can be observed. 1) From Mondays to Thursdays, the average numbers of trips per

person-day were similar. 2) More trips were observed on Fridays. 3) Weekends had fewer numbers of trips, with Sunday having the least. Additionally, fewer trips were made on holidays (e.g., Memorial Day).

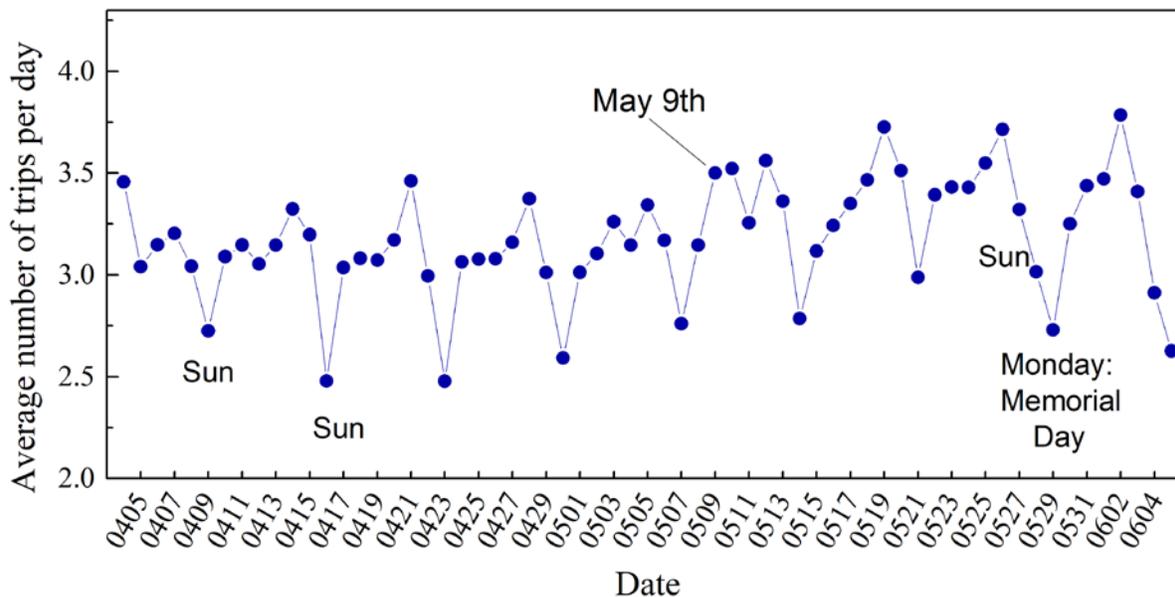


Figure 21. Graph. Weekly trip rate pattern

3.3.2 Trip Departure Times Distribution

Figure 22 compares the distribution of departure times inferred from the app-based data and that obtained from the survey data. The PSRC distribution showed two peaks (8:00 to 10:00, 16:00-18:00), corresponding to morning and afternoon peak commute periods. The distribution from the app-based data showed that both morning and evening peaks were greatly mitigated. On the other hand, the weekend distribution for the app-based data was single-peaked and extended into the afternoon. The arrival time distributions are not presented in this report, as they were similar to the departure time distributions.

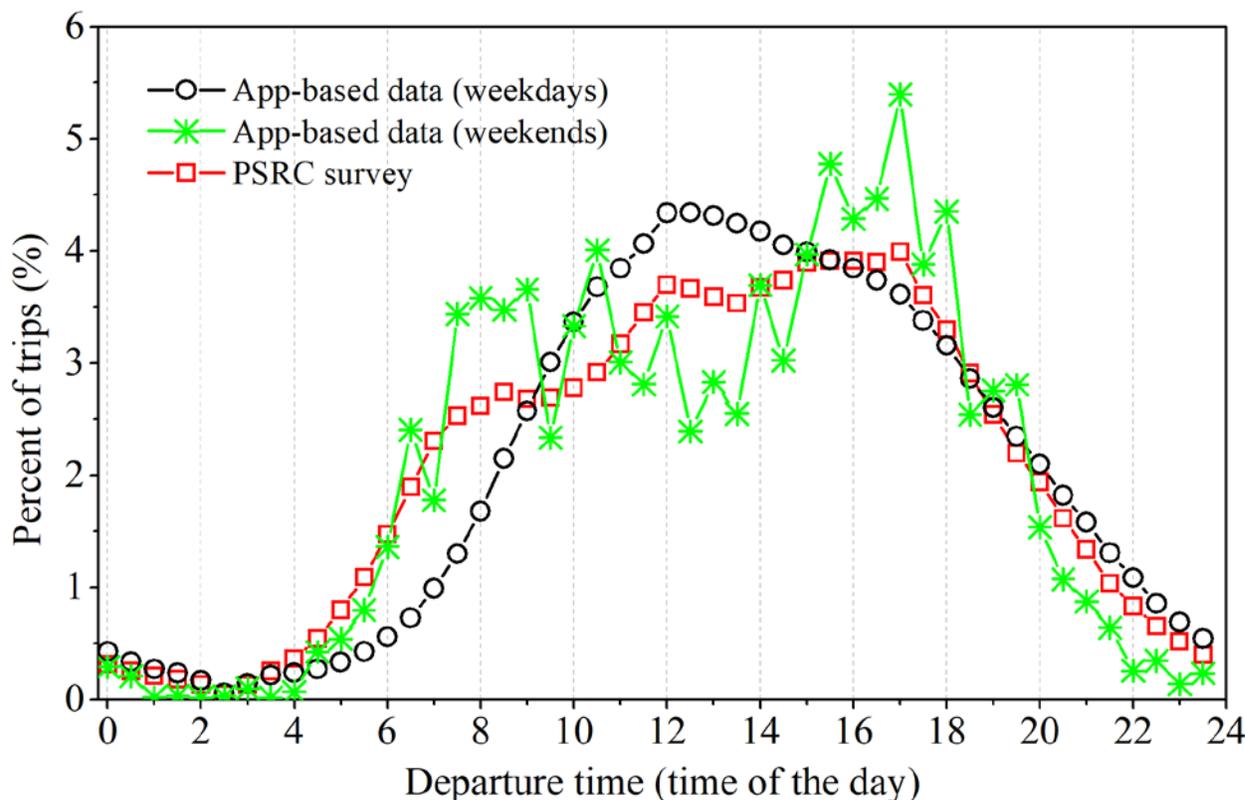


Figure 22. Graph. Departure time distribution

3.3.3 Trip Length and Travel Time

Figure 23 shows the distributions of trip distances (in kilometers) for the app-based and PSRC survey data. The two are quite consistent with each other despite the much larger variance for the PSRC curve because of its small sample size. This figure again shows that there was more over-representation of short trips (of less than 500 meters) in the apps data than in the survey data. In addition to the reasons discussed previously (e.g., under-reporting of short-trips in the survey data; or some signaling noises in the apps data that may have been mistakenly identified as trips), the issue could also have been due to the fact that trip distances were calculated as Euclidean distances in this study, which typically are shorter than real-world trip distances in a road network (e.g., those recorded in the survey data).

From the cumulative curve (Figure 24), it can be observed that about 70 percent of trips were shorter than 10 kilometers, while nearly 50 percent of trips were between 1 and 10 kilometers. Distributions on weekends are not presented, as they were not distinguishable from weekdays.

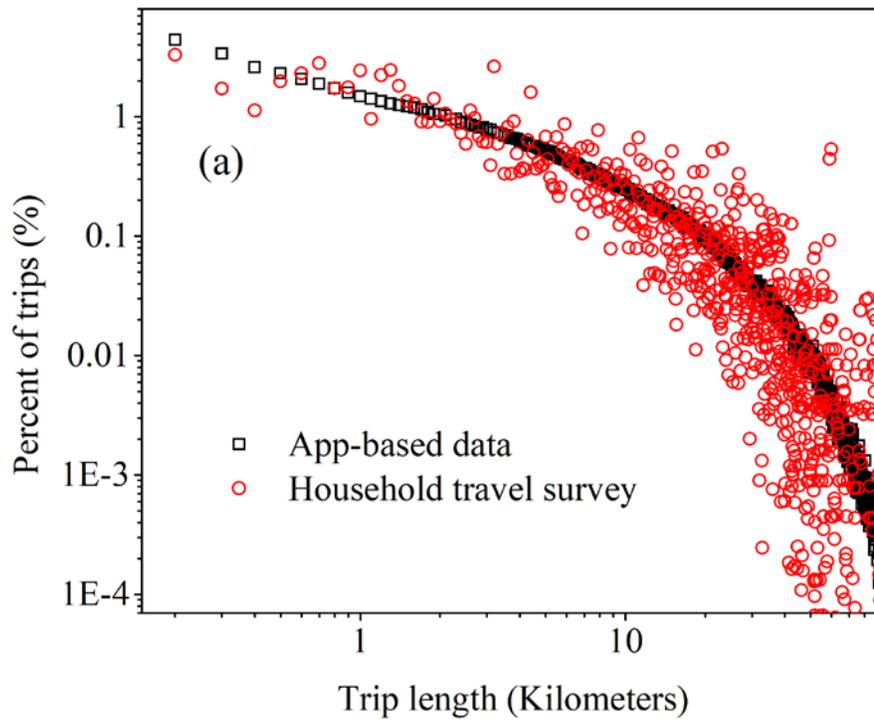


Figure 23. Graph. Distribution of travel distance

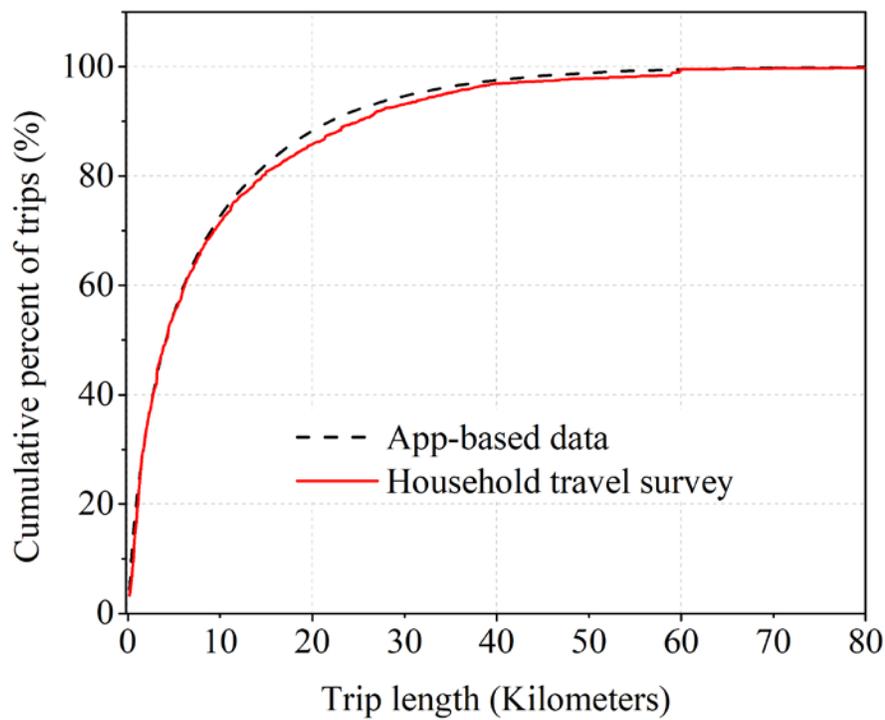


Figure 24. Graph. Cumulative distribution of travel distance

Figure 25 shows the distribution of trip times. Short travel times (of less than 25 minutes) tended to be under-represented, and longer travel times (longer than 25 minutes) tended to be more over-represented in the apps data than in the survey data. The difference is more pronounced in Figure 26. This was consistent with the results of activity durations and trip rates, which can be explained by Figure 16.

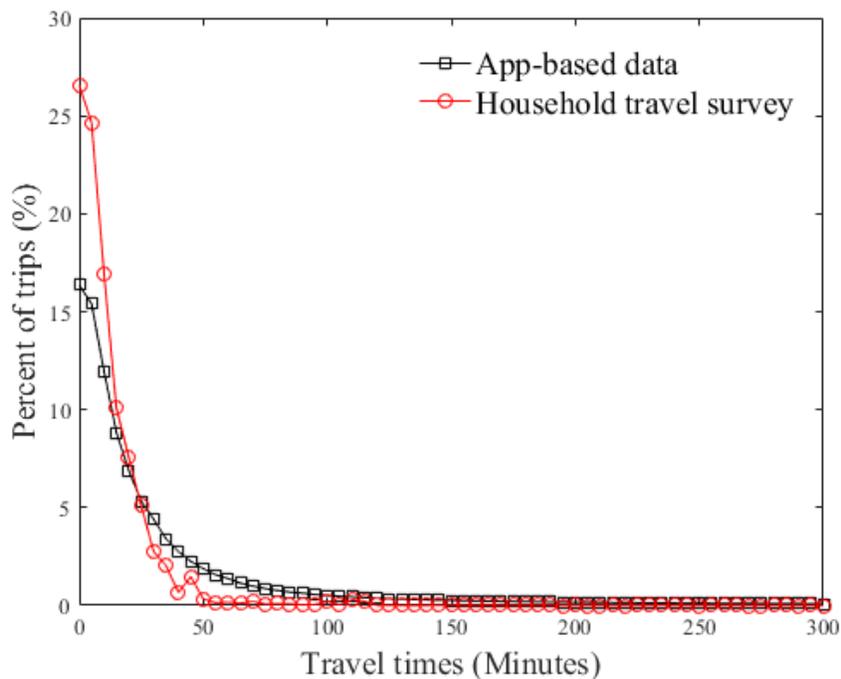


Figure 25. Graph. Distribution of travel times

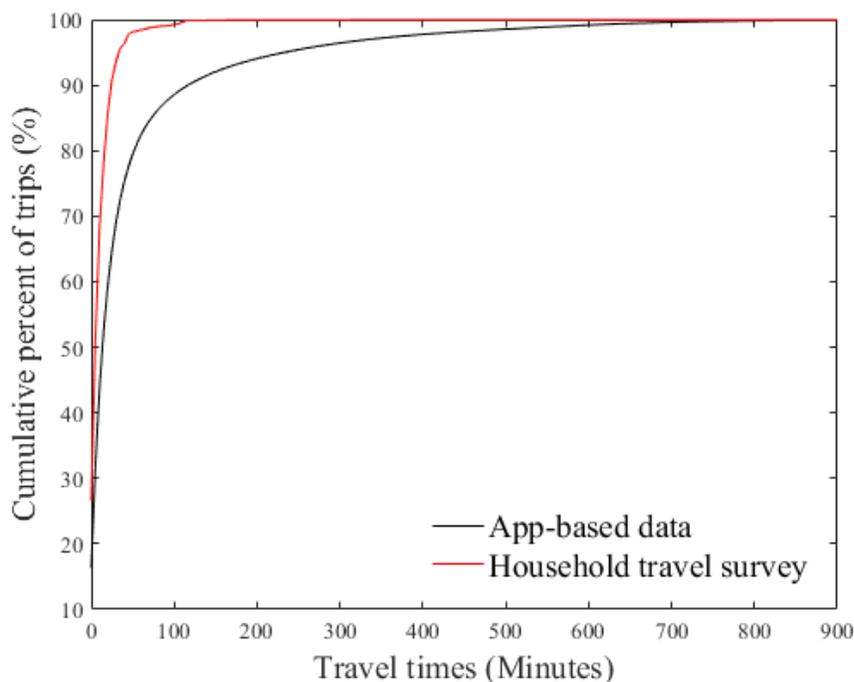


Figure 26. Graph. Cumulative distribution of travel times

3.3.4 Estimated travel demand

When individual trips are aggregated to the zone level, then origin-destination patterns can be analyzed. Note that the OD trips observed directly in the app-based data were not comparable with the metropolitan planning organization (MPO) OD trips, as the latter attempts to capture all trips in a region, whereas the former represented only users who appeared in the data.

Therefore, a scale-up OD estimation method was developed to estimate the OD demand for all trips from the app-based data (see Appendix B). The MPO OD matrix used in this research was obtained from SoundCast (2014 base year), which is a travel demand model built for the Puget Sound region. This OD matrix contained only the internal trips (trips completed within the PSR) by residents.

Figure 27 shows the spatial distribution of trip origins estimated from the app-based data and SoundCast. The traffic analysis zones (TAZs) generating a large number of trips were concentrated in several specific areas: major university campuses (University of Washington), major airports (Sea-Tac International Airport), downtown areas (Seattle, Bellevue), and major high-tech campuses (South Lake Union – Amazon Campus). In comparison to the PSRC model results (Figure 27-b), a majority of zones with larger numbers of generated trips correlated with the app-based data estimation results, suggesting that the app-based data were able to capture regional travel demand to some degree. The spatial distribution of trip destinations was similar to that of trip origins and therefore is not presented in this report.

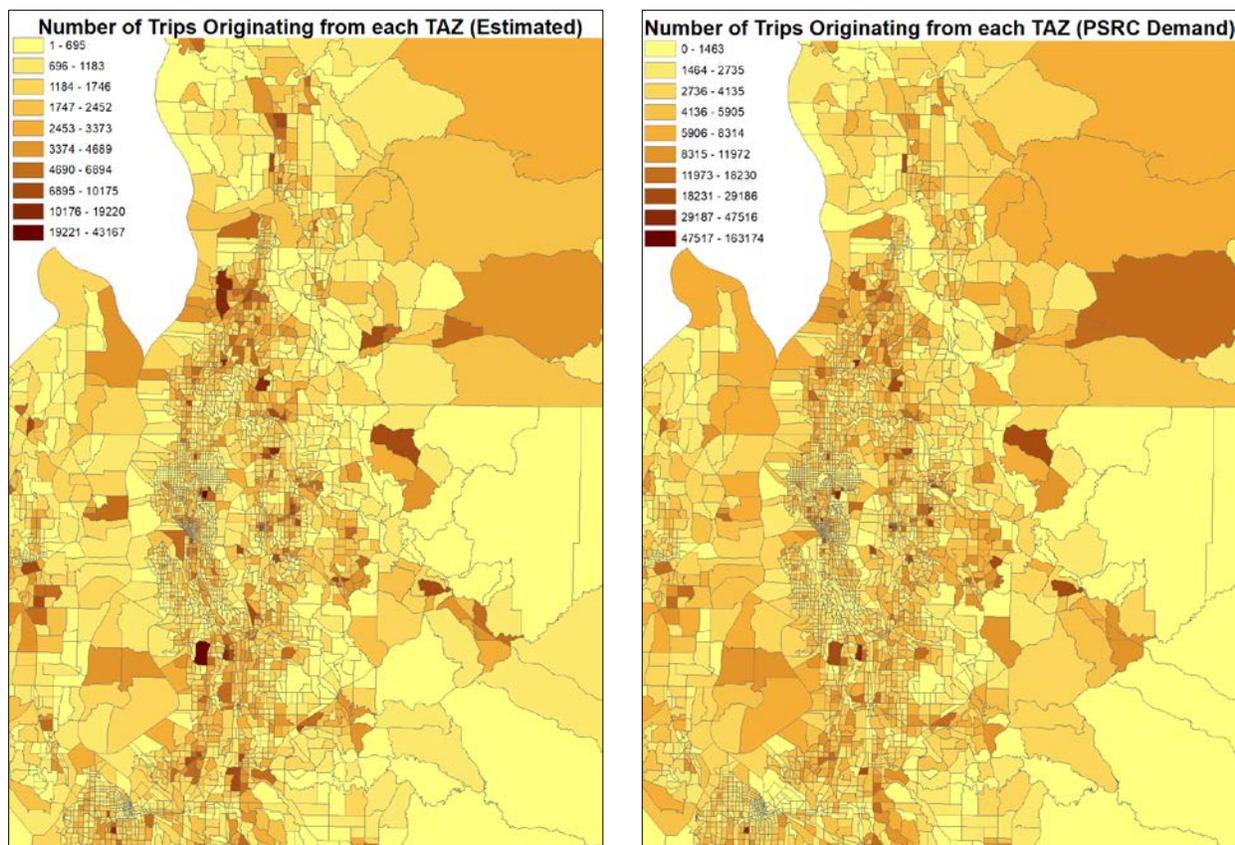


Figure 27. Map. Spatial distribution of trip origins. (a) Estimated results from app-based data, (b) SoundCast results

The correlation between estimated values and SoundCast results at the TAZ level is shown in Figure 28. Overall, the linear regression between estimated trips and MPO model trips had an R-squared value of 0.5049, which is slightly lower than the results generated by vehicular GPS data (0.588) (Chen et al., 2017;). Similar observations were found from the correlation for the trips heading to TAZs and are therefore not presented in this report.

Figure 29 shows the correlations between estimated OD demands and PSRC OD demands, which is lower (0.3636) than if only origins were compared (0.5049). This suggests that one should be cautious when directly using the app-based data for capturing a region's origin-destination travel patterns.

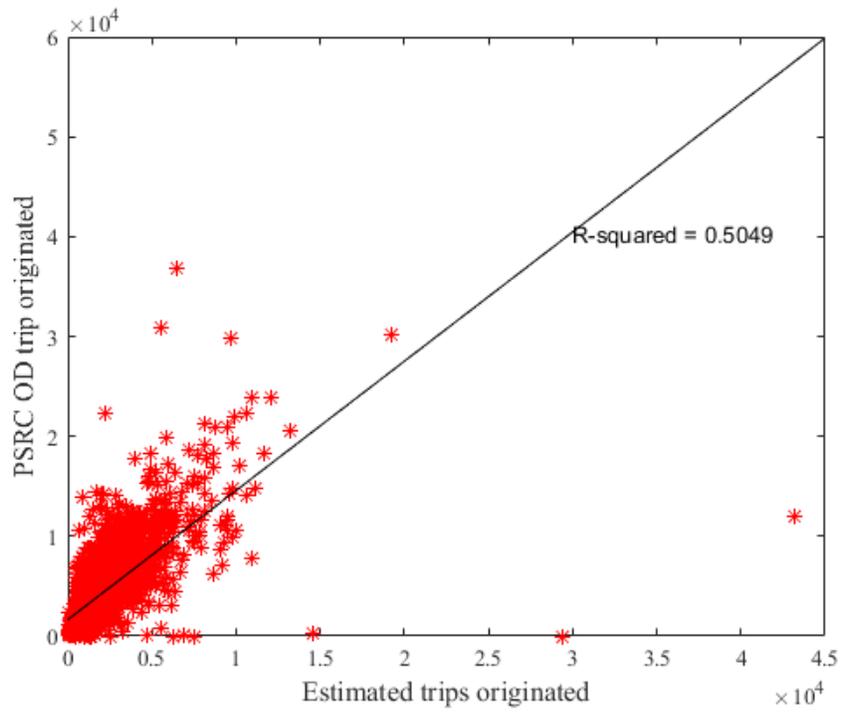


Figure 28. Graph. Correlations between estimated trip origins and MPO trip origins

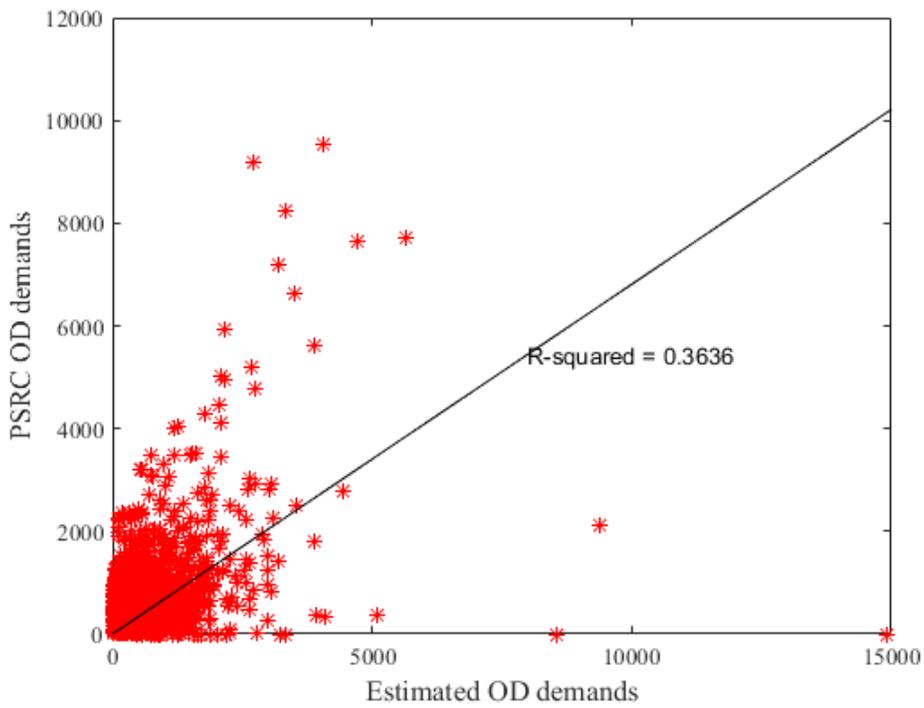


Figure 29. Graph. Correlations between estimated OD demands and PSRC OD demands

3.4 May 9th Data Shift

As mentioned in Section 3.1.3, a sudden change in the data set occurred starting on May 9th. We called this phenomenon the “May 9th data shift.” One or more factors may have contributed to this phenomenon: (1) As the data were from various apps, the data provider may have obtained access to more apps on May 9th. (2) The number of available apps did not change, but the number of users contributing to the data may have increased. (3) Neither the number of apps nor the number of users may have changed, but apps may have enhanced their services by requesting the locations of their users more frequently. (4) Neither the number of apps nor the number of users may have changed, but usage patterns may have changed (i.e., users started using apps more frequently). The fourth possibility may be quickly dismissed, as it is unlikely that all users collectively changed their usage patterns within a day.

Figure 30 shows the temporal evolution for the daily number of unique IDs. A slight 3 percent increase can be observed on May 9th, in comparison to the previous day, May 8th. The number of observations per ID by day is shown in Figure 31. In comparison to the previous day, the number of observations per ID increased by 33 percent and remained at this level for the rest of the study period. Figure 30 and Figure 31 together suggest that the increase in data size starting on May 9th was most likely attributed to the data provider obtaining more observations from each device. We later confirmed with the data provider that on May 9th, more app developers signed up for its SDK, resulting in more apps contributing to its data collection (per device). This indicates that big data, such as the app data studied here, are dynamic and constantly changing. As a result, their data properties also need to be investigated periodically; see the discussion section for more details.

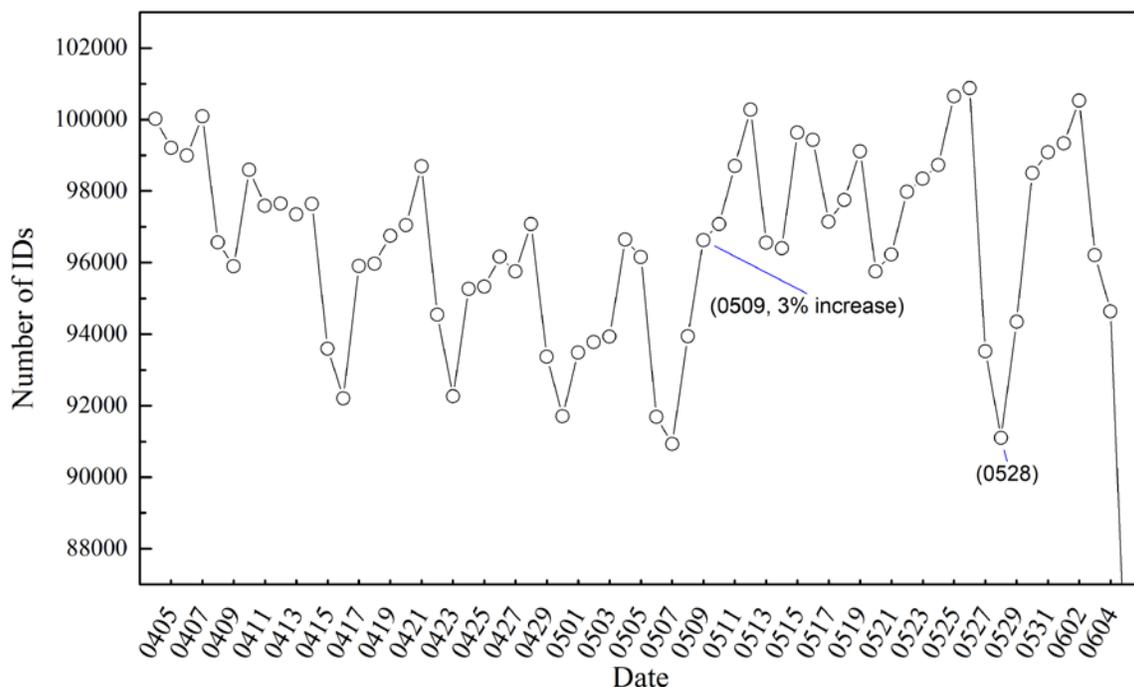


Figure 30. Graph. Evolution of daily number of unique IDs (zeroth order)

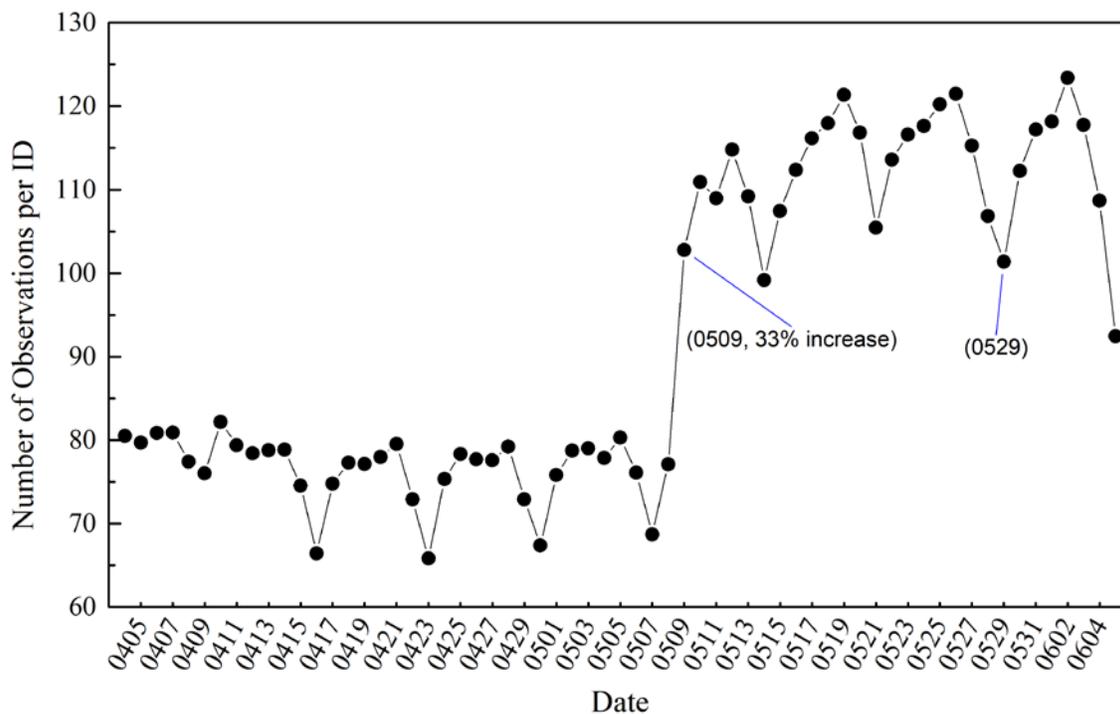


Figure 31. Graph. Evolution of daily number of observations per ID (zeroth order)

Figure 32 shows the temporal evolution of location accuracy, using three statistics: the 1st, 2nd and 3rd quartile. While the 2nd quartile showed no change before or after May 9th, the 1st quartile decreased and the 3rd quartile increased. Because location accuracy is linked to the technologies used for data collection, the broader distribution also suggests an increased variety of the data in terms of location accuracy after May 9th. The implications are discussed in Section 3.5.

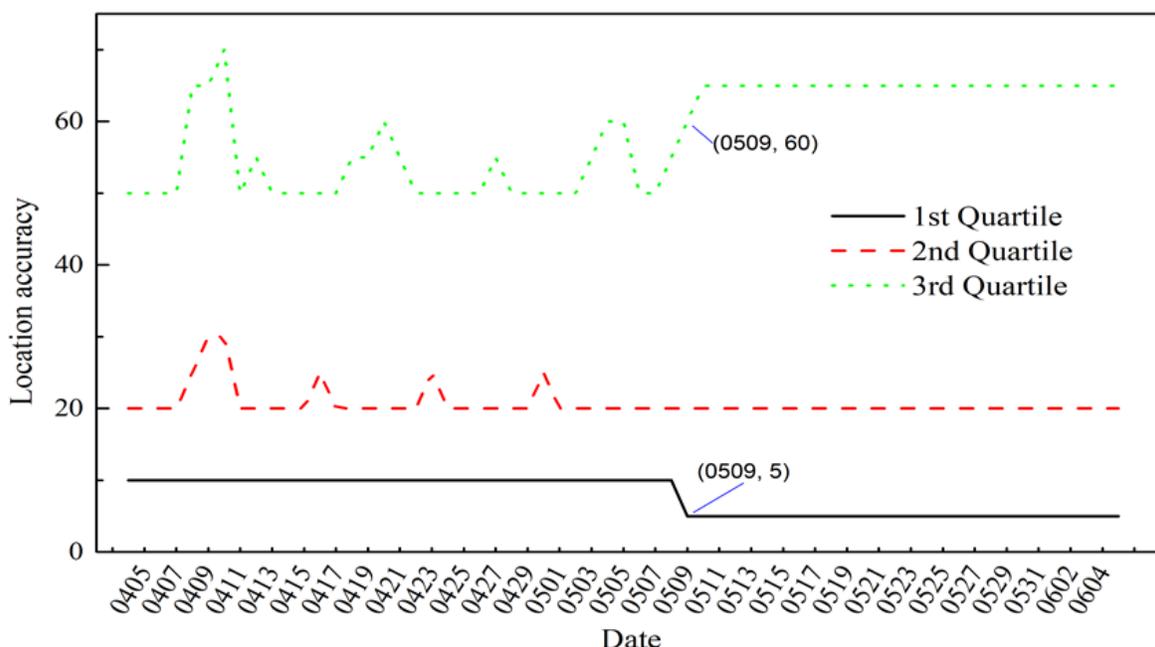


Figure 32. Graph. Evolution of location accuracy (zeroth order)

Figure 33 shows reduced time intervals between two consecutive observations (i.e., denser observations) after May 9th. Conversely, Figure 34 shows the evolution of temporal sparsity, which is represented by the number of time slots (30 minutes for each slot) that had at least one observation. **The slight increase in the temporal sparsity suggests that a significant increase in data size did not necessarily significantly reduce the temporal sparsity of the data.**

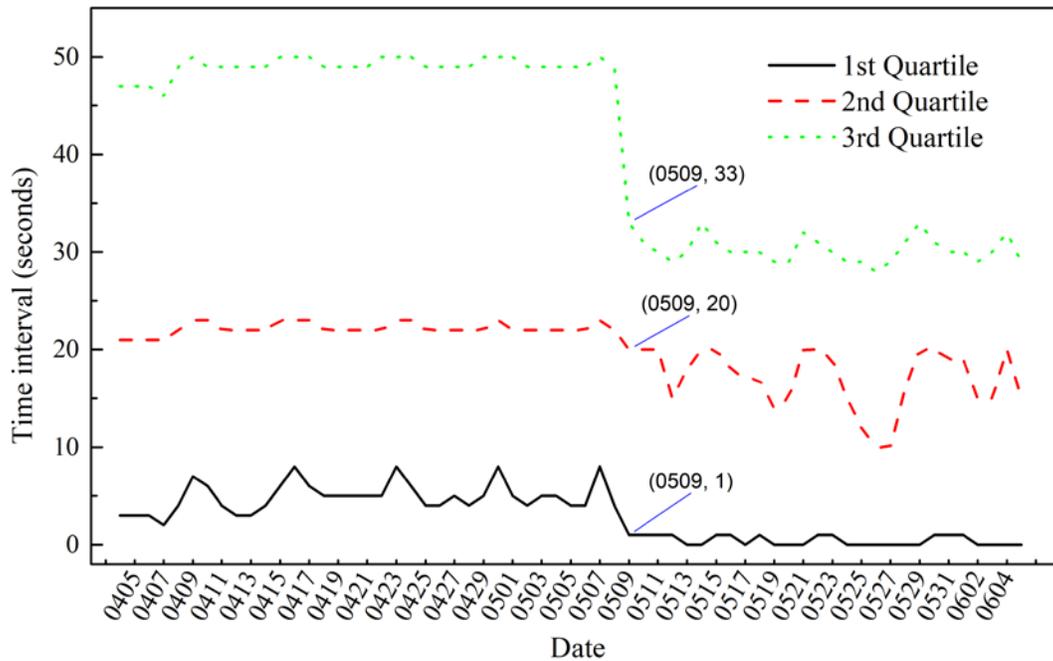


Figure 33. Graph. Evolution of time interval (zeroth order)

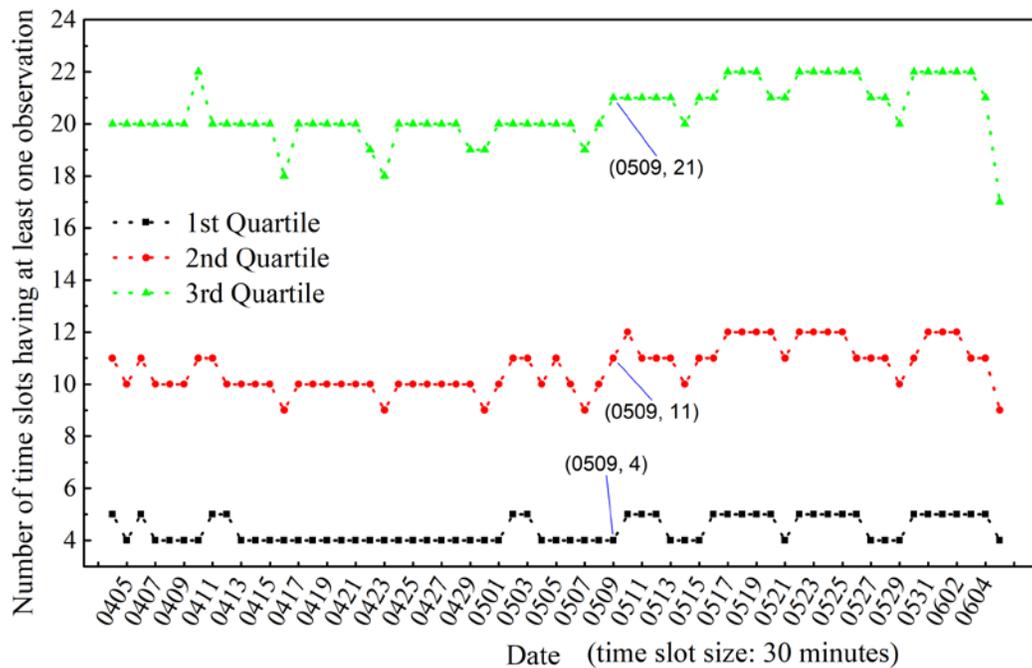


Figure 34. Graph. Evolution of temporal sparsity (zeroth order)

The rest of this section describes more analyses conducted to examine how the increase in data size influenced first- and second-order properties.

As shown in Figure 35, no clear difference could be observed between activity durations (first-order property) before and after May 9th. However, the mean trip rate (second-order property) increased from 3.11 before May 9th to 3.37 after May 9th, as shown in Figure 36. The figure also shows that after May 9th, the trip rate distribution showed a pattern more similar to the PSRC trip rate distribution, e.g., the percentage of higher trip rates increased, and the percentage of lower trip rates decreased. This could be due to the fact that some missing trips in the *before* data were now revealed in the *after* data with more observations. The improvements in trip rates after May 9th, however, was not significant, given the significant increase in the number of observations (per ID) in the *after* data set.

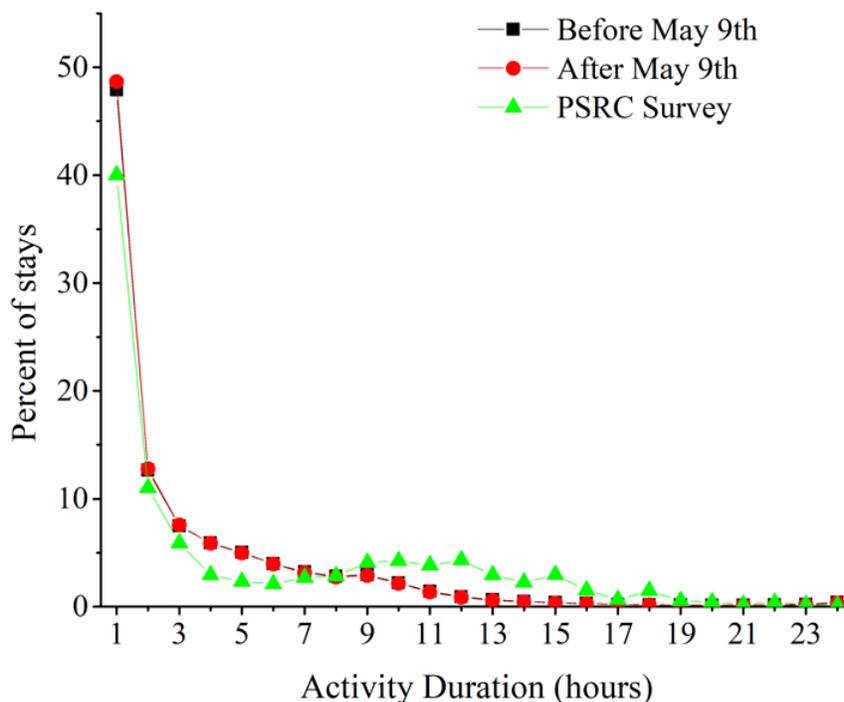


Figure 35. Graph. Comparison of activity duration (first^t order) before and after May 9th

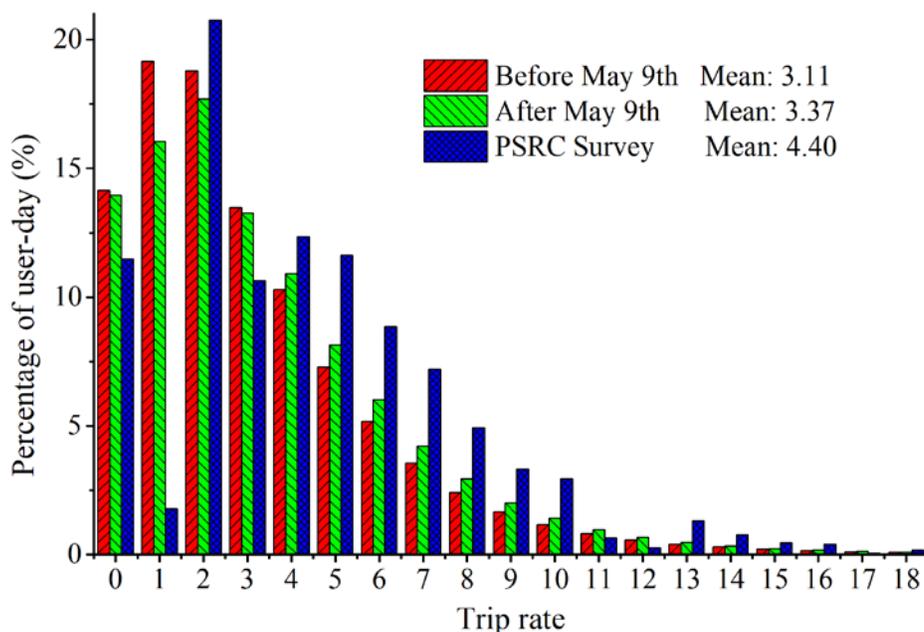


Figure 36. Graph. Comparison of trip rate (second order) before and after May 9th

Figure 37, Figure 38, and Figure 39 compare the distributions of departure time, trip length, and travel time before and after May 9th, respectively. No major differences were identified except for a slight difference in the distribution of travel times. The distribution of travel times for after May 9th was closer to that calculated from the PSRC survey data. Again, similar to the trip rate distribution in Figure 36, the improvement was fairly marginal.

In summary, we conclude that after May 9th, the average observations per device increased substantially (about 33 percent, which however led to either unchanged or only minor improvements in properties (especially the first- and second-order properties). For this reason, the data properties presented in Sections 3.1 – 3.3 were calculated from the entire data set.

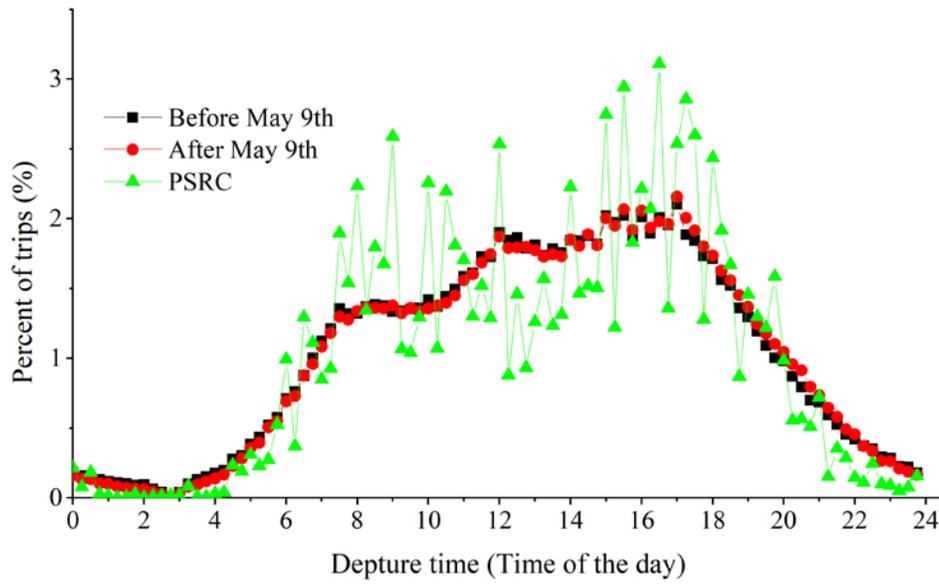


Figure 37. Graph. Comparison of departure time distributions (second order) before and after May 9th

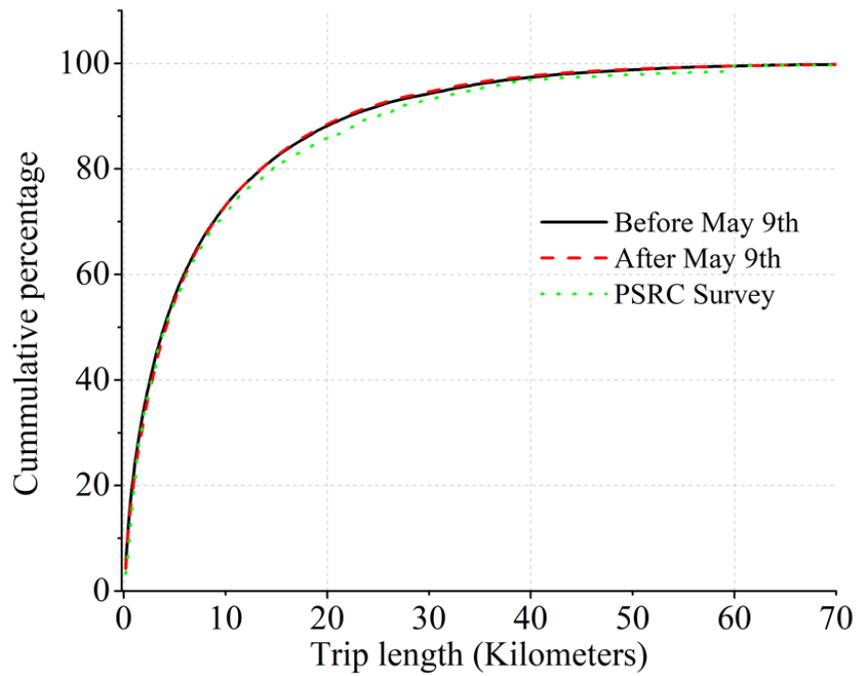


Figure 38. Graph. Comparison of cumulative distributions of trip length (second order) before and after May 9th

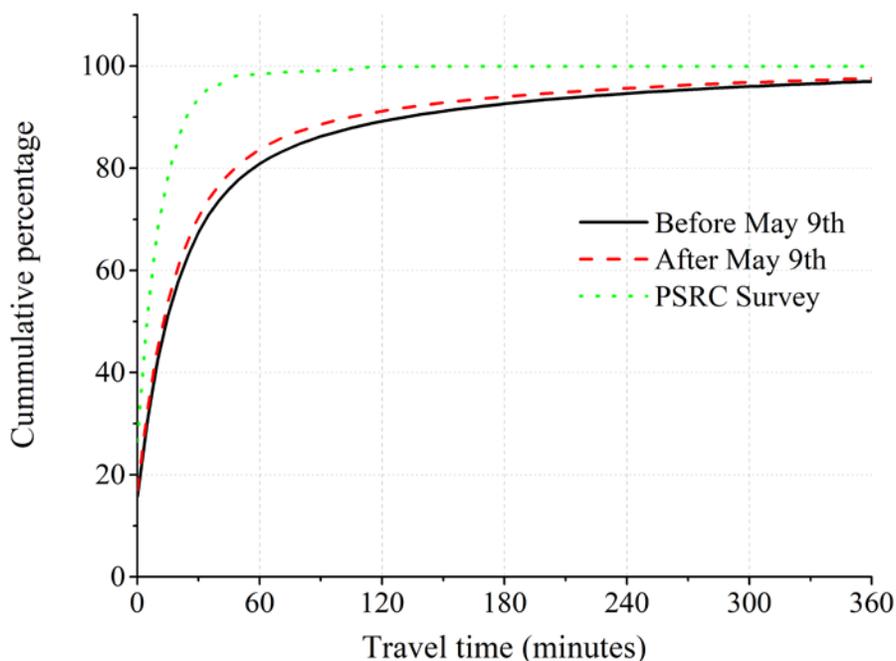


Figure 39. Graph. Comparison of cumulative distributions of travel time (second order) before and after May 9th

3.5 Summary

Connecting to our previous report (Chen et al., 2017), this section provides a summary of the major findings based on the five data sets (big and small) that we have analyzed. The five datasets included the following:

- 2002 mobile phone data for the Buffalo-Niagara region;
- corresponding household travel survey data for the Buffalo-Niagara region;
- app-based data for the Puget Sound region;
- the corresponding Puget Sound household travel survey data; and
- vehicular GPS data for part of the Seattle downtown area.

Table 3 gives an overview of these data sets⁷.

⁷ The vehicular GPS data were for a small area of the City of Seattle. The analysis results were not very applicable to travel patterns for the region. Therefore Table 3 does not include this data set.

Table 3. Summary of datasets

Datasets	Study period	Study area	Number of samples	Sampling rate
App-based data	April 4 th - June 5 th 2017	Puget Sound Region (Washington State, US)	462,401 unique IDs	11.9% (# of people sampled divided by the population, 2015 PSRC)
Mobile phone data	April 2014	Buffalo-Niagara Region (New York State, US)	933,508 users	82.2% (# of people sampled divided by the population, 2010 Census)
2017 PSRC household travel survey data	April-June 2017	Puget Sound Region (Washington State, US)	3,277 households; 6,235 persons	0.16% (# of people sampled divided by the population, 2017 PSRC)
2002 Buffalo-Niagara regional transportation survey data	2002	Buffalo-Niagara Region (New York State, US)	2,779 households; 6,636 persons	0.59% (# of people sampled divided by the population, 2000 Census)

We can further summarize the characteristics of the five data sets (also see Table 4) as follows:

- Big data: passively generated data from billing, app usage, and other primary purposes that can be used for transportation planning applications (Chen et al., 2016); large data size covers a significant portion of a population in a region.
 - Mobile phone data: traces generated as users make phone calls and send/receive text messages; single-sourced positioning technology relying on triangulation of cellular towers, with an accuracy level ranging from a few hundreds to thousands of meters (Calabrese et al., 2011; Chen et al., 2016; Iqbal et al., 2014).
 - Vehicle GPS data: traces generated as vehicles (trucks and passenger cars) equipped with GPS operate; single-sourced positioning technology relying on GPS, with an accuracy level of a few meters (Chen et al., 2017).
 - Apps data: traces generated as users use various apps on smart phones; multi-sourced positioning technology relying on GPS, Wi-Fi, and cellular towers.
- Small data: actively solicited from targeted participants through a probabilistic sampling process in a well-defined target population; small sample sizes ranging from 0.5 percent to 1 percent.
 - Buffalo-Niagara household travel survey data: respondents were asked to report all trips and associated attributes (origins and destinations in exact addresses or closest intersections, departure and arrival times, mode of transportation, travel times, etc.); single-source positioning technology relying on user-reported addresses that were then translated into geo-coordinates.

- Puget Sound region household travel survey data: respondents were asked to report all trips and associated attributes (origins and destinations in exact addresses or closest intersections, departure and arrival times, mode of transportation, travel times, etc.); single-source positioning technology relying on user-reported addresses that were then translated into geo-coordinates⁸.

As the two small household travel survey data were similar in nature, we grouped them into one type, together with the three types of big data. These four types are presented in Table 4, which summarizes their unique, key characteristics in pro and con categories.

Table 4. Pros/cons of different data

Data type	Pros	Cons
App-based data	Large size; higher observational frequency; mixed GPS, WiFi, Bluetooth, and cellular tower positioning; presence of trace information; inexpensive; continuous	Non-probabilistic samples; missing trips/activities; no demographics information; unknown underlying population
Mobile phone data	Large size; presence of trace information; inexpensive; continuous	Lower observational frequency; lower positioning accuracy; Non-probabilistic samples; missing trips/activities; no demographics information; unknown underlying population
Vehicular GPS data	Large size; higher observational frequency; high positioning accuracy; presence of trace information; continuous	Non-probabilistic samples; missing trips/activities; no demographics information; unknown underlying population; only for vehicular travels
Household travel survey data	Probabilistic but small samples; designed to be representative; rich information on activity and travel patterns; with demographics and attitudinal information	Lack of trace information on routes; Lack of information for non-residents; expensive; infrequent data collection; static information

Clearly, small data are often collected via a rigorously designed sampling/collection process targeting a specific population. In other words, they are designed to be representative of the underlying population. Big data, however, are mostly the by-product of certain primary

⁸ There was a separate, much smaller, GPS-only sample in which respondents were asked to download a GPS app onto their phones, and the app recorded all traces and interacted with respondents to verify trip-related attributes. At the time of the writing of this report, this sample had not yet been obtained and therefore was not included in this report. It may be analyzed in future phases.

purposes, which usually do not follow any well-designed data collection process, and hence the collected data are often not representative. On the other hand, small data are often static (i.e., collected every 5 to 10 years), cover a tiny fraction of the underlying population, and lack trace information on routes, whereas big data surpass the small survey data with their volume and continuity, and can contain more complete information on short trips and routes that are often neglected in survey data. Another unique feature of big data is that they are good at showing “what happened” but not “why that happened,” whereas small data are behaviorally much richer and can help explain the fundamental reasons underlying observed travel phenomena.

The above differences between big data and small data contribute to the overall properties of the data (i.e., zeroth order) and the travel-related metrics derived from the data (i.e., first and second order). Below we first summarize the similarities and differences across the data sets regarding the zeroth order. Again the results from the vehicular GPS data are not included here because of the limited study area.

- *Sampling rate.* Both big data sets (i.e., mobile phone data and app-based data) had a high sampling rate in comparison to the region’s population (e.g., 11.9 percent for app-based data and 82.2 percent for mobile phone data), while the sampling rates for the survey data were much smaller (0.16 percent for the PSRC travel survey and 0.59 percent for the Buffalo travel survey). Note that the “sampling rates” for big data are not statistical sampling rates; rather, they should be interpreted as some form of “market penetration” of the devices. It is also important to note that the underlying populations for big data are likely very different from the resident populations for household travel surveys; the vast majority of users in our big data were observed only for very short periods, suggesting that they may not have been residents or may have been residents but not actively using their mobile phones or mobile apps.
- *Intra-day temporal sparsity.* As shown in Figure 40, on weekdays, the app-based data had three small peaks within a day, whereas the mobile phone data showed two peaks. From midnight to early morning, the fraction of IDs in the app-based data was much larger than that of the mobile phone data. This is inherently related to the underlying data generation process: app-based data are derived from apps usage whereas mobile phone data are from phone calls and text messages. During the night and early morning, the number of phone calls and text messages largely subsided while app use was still substantial (around 25 percent). In addition, one can observe that most usages in the app-based data and mobile phone data (either weekdays or weekend) occurred between 9:00 AM and 8:00 PM, whereas traffic usually has distinct early morning and afternoon peaks. This indicates that if the derived patterns from big data are used for travel pattern analysis, one needs to be cautious because mobile apps or mobile phone data show only device usage, not necessarily user travel intensity. It is clear that the temporal sparsity feature of big data has direct impact on the accuracy of the activity locations (stays) derived from big data, and the resulting travel patterns estimated from the data (see Figure 16).

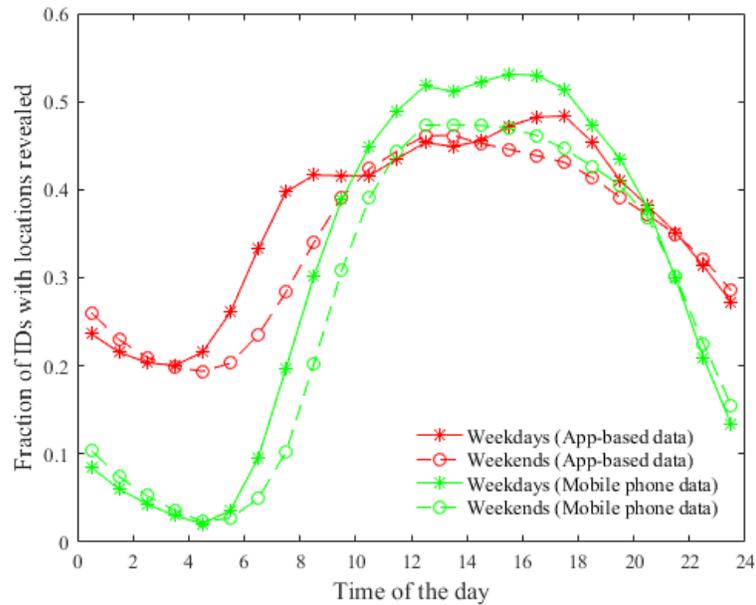


Figure 40. Graph. Fraction of IDs with their locations revealed at time of a day (both big data sets)

- *Weekly pattern of the number of observations.* Both big data sets showed a consistent weekly pattern, as shown in Figure 41, suggesting that both data sets were good candidates for weekly trend analysis. Generally, weekdays had more observations than weekends, and Sundays had the least. In addition, the drops in phone calls were slightly more severe than those for app use. The drops of the two curves imply that on weekends people tended to make fewer phone calls, send fewer messages, and make less use of mobile apps.

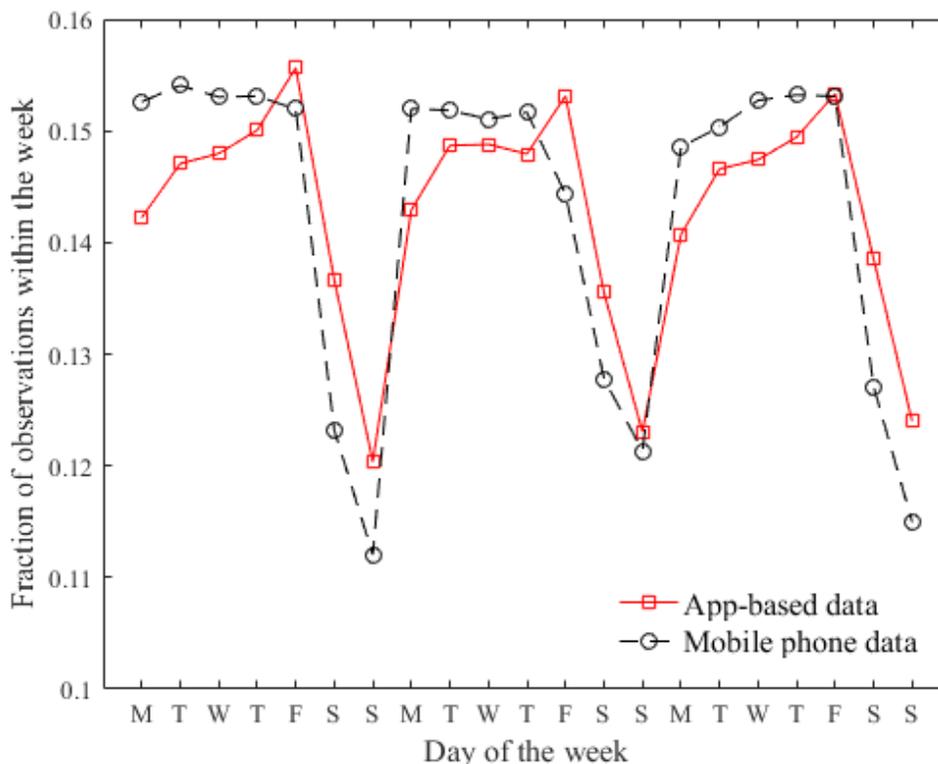


Figure 41. Graph. Fraction of observations within the week (both big data sets)

- *Location accuracy.* The location accuracy of the app-based data was generally much higher than that of the mobile phone data, according to the cumulative distribution of data location accuracy (see Figure 42). For example, the 85th percentile of location accuracy for the app-based data was 100 meters, while this statistic for the mobile phone data was about 700 meters. From 700 to 2400 meters, the two curves increase at a similar rate, indicating that both data sets had similar distributions for data accuracy in the range of 700 to 2400. This is probably due to the fact that both data sets were the produce of similar positioning technology at this accuracy range (e.g., cellular tower technology).

The overall properties of big data, especially the temporal sparsity and location accuracy, largely determine how accurate the identified activity locations from the data can be in comparison to ground truth or certain benchmarks (e.g., those from travel surveys). As shown in Figure 16, because of temporal sparsity and inaccurate location information, certain activity locations may be missed or the arrival/departure times may be incorrectly identified. This leads to underestimation of activity durations and trip rates and overestimation of trip travel times. As for OD demands, because big data are not expected to represent the underlying population well, the correlation between big data-derived OD demands and the benchmark OD demands are mediocre in most cases. More detailed comparisons regarding the first- and second-order properties are summarized as follows:

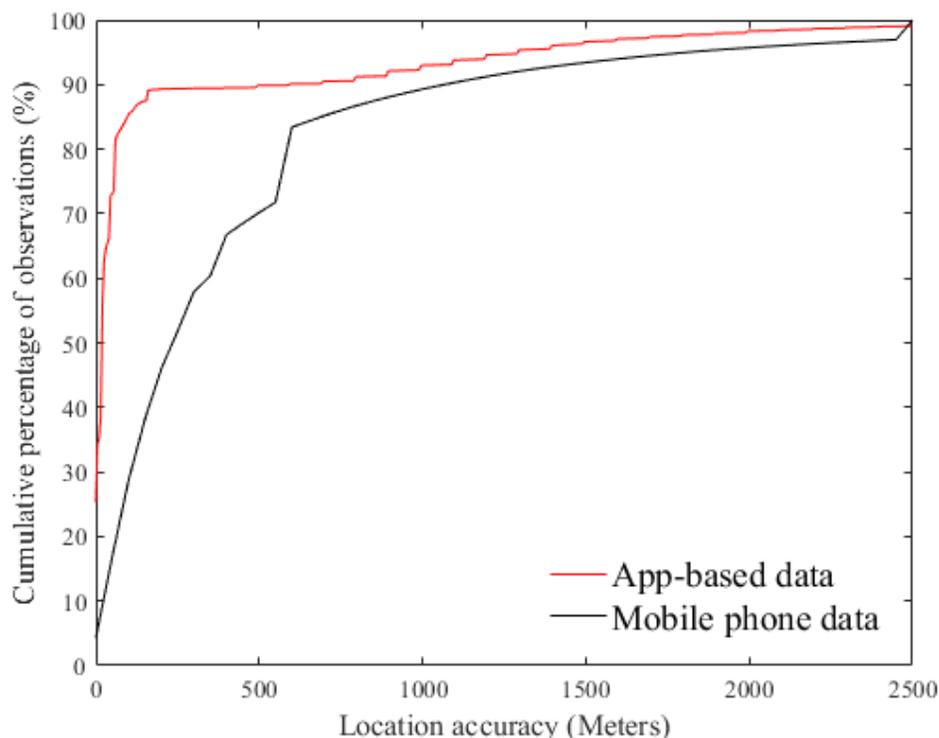


Figure 42. Graph. Cumulative distribution of data location accuracy (both big data sets)

- *Activity duration.* The app-based data set yielded about 48 percent of the durations of less than one hour, corresponding to about 40 percent from the PSRC travel survey. The reverse seemed to be true for the mobile phone data: about 32 percent of activities were identified from the mobile phone data lasting less than one hour, corresponding to 40 percent from the Buffalo travel survey. In addition, there seemed to be an over-estimation of activities lasting between 2 and 6 hours for the app-based data and an underestimation of activities lasting between 7 and 14 hours for the mobile phone data in comparison to their corresponding survey data. The deviations of activity durations estimated from app-based and mobile phone data sets may be attributed to several factors, including the temporal sparsity of the two data sets, representativeness issues of the data (i.e., only the data from specific population were collected), and the fact that big data sources can help identify short trips (e.g., a trip to a coffee shop near the work place) that are typically not reported in travel survey data.
- *Home census tract correlation.* Correlation between the number of inferred residents from both big data sets and the population statistics at the census tract level were 0.91 for the app-based data and 0.43 for the mobile phone data. The higher correlation of app-based data demonstrates their better capability to infer home census tracts which is crucial for regional travel pattern analysis, than mobile phone data. The difference between the two may be due to the following two reasons: 1) the generally lower accuracy of mobile phone data makes it harder to correctly identify home census tract; and 2) app-based data is also temporally less sparse than mobile phone data.

- *Trip rate.* For both big data sets, the estimated trip rates were significantly lower than those obtained from survey data. For the app-based data, the estimated mean trip rates (per day) were 3.23 for weekdays in comparison to 4.40 from the PSRC survey data. For the mobile phone data, the estimated mean trip rates (per day) were about 1.78 for weekdays in comparison to 3.89 from the Buffalo travel survey data for weekdays. Both big data sets experienced under-estimation bias, although to a lesser degree for the app-based data. Note that although survey data are not “ground-truth,” they are believed to have less significant bias issues than big data because of their distinct, controlled data collection process. Hence they were used as a benchmark for comparison.
- *Departure time.* Both surveys clearly showed three peaks in the morning at 8:00 AM, at noon, and in the afternoon between 3:00 and 5:00 PM, reflecting peaks in travel patterns (morning commute peaks, noon for lunch breaks, and afternoon commute peaks). For both big data sets, they were consistent with the surveys only in the afternoon peak.
- *Travel time.* The cumulative percentage of trips from both surveys shared similar patterns in terms of travel times, as shown in Figure 43. For shorter travel times (0-100 minutes), the two survey curves grow much faster than those of the app-based data and mobile phone data, indicating that a much smaller percentage of users in the survey data had long travel times, or survey participants tended to ignore/forget shorter trips when participating in the survey, or both. Substantially more short-time trips were captured by the app-based data than by the mobile phone data. This finding is consistent with much the lower trip rate derived from the mobile phone data than the app-based data and the higher levels of temporal sparsity observed for the mobile phone data.

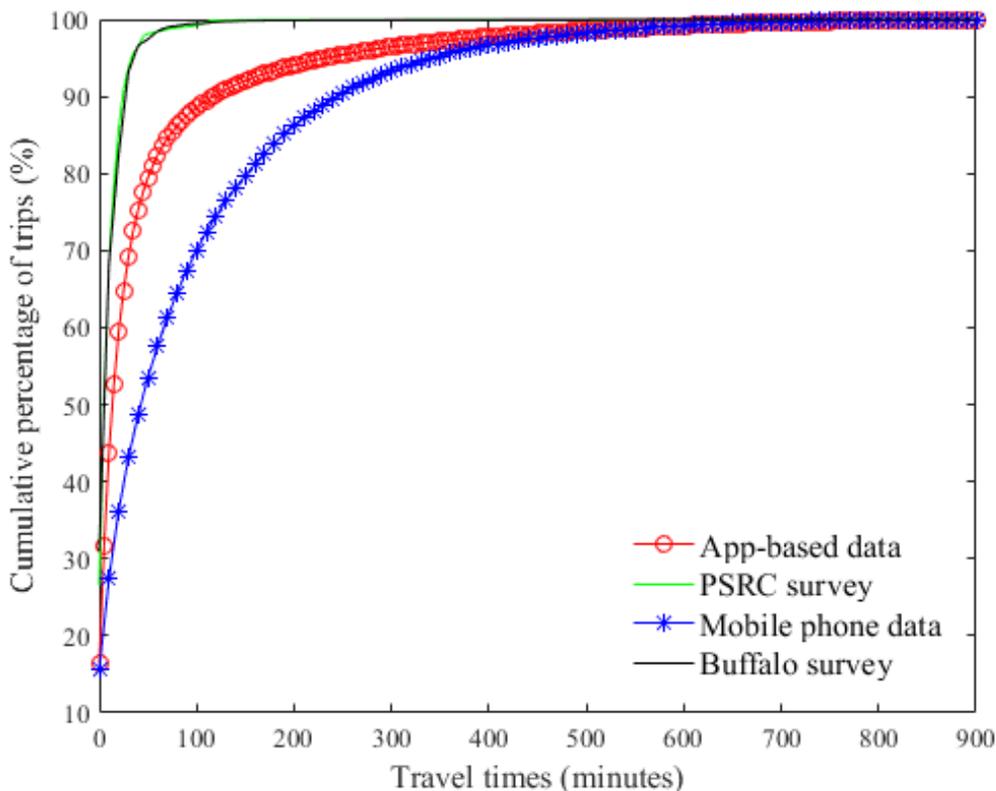


Figure 43. Graph. Cumulative distribution of travel times (four data sets)

- *Correlation between MPO OD and estimated OD.* The correlation between the estimated OD of the app-based data (0.36) and the MPO OD was less than that for the mobile phone data (0.66) (Chen et al., 2017). Both are considered low, indicating that the estimated OD demands from big data do not sufficiently represent MPO demand matrices. Therefore, it is not advisable to use OD demand matrices developed directly from big data sources. 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 big data sources.

3.6 Discussion

Within the short, one-year period from the last study (Chen et al., 2017), during which mobile phone and vehicle GPS data were studied, to this study in which apps data were studied, the technologies used for data generation evolved in at least two aspects. First, the scope of user activities greatly expanded from just phone calls and text messages for mobile phone data to the use of many different apps for app-based data. Second, the positioning technology also evolved from being single-sourced (e.g., either cellular towers or GPS) for mobile phone or vehicle GPS data to multi-sourced (cellular towers, GPS, WiFi, and Bluetooth) for the apps data. Consequently, as shown in our last and this report, both spatial and temporal properties associated with the data, in particular, locational accuracy and temporal sparsity, improved.

Changes in these properties allowed us to connect to the underlying population better and to capture travel patterns in a more accurate and complete way, as reflected by the high correlation between inferred residents and the census population density (0.91 in Figure 14), an estimated average trip rate that was much closer to that of the household travel surveys (3.23 from apps data vs 4.40 from PSRC travel survey data, in comparison to 1.78 from mobile phone data vs 3.89 from Buffalo travel survey data), and the identification of both morning and afternoon peaks in the apps data (although delayed) in comparison to a single afternoon peak shown in the mobile phone data. Despite these improvements, however, the analyses in this report also showed that significant discrepancies still exist in comparing the travel patterns (e.g., trip rates, OD demands, etc.) estimated from big data (e.g., app-based data) and those from other, potentially more representative, data sources (such as survey data, as shown in Table 4).

It is clear that the changes in the technologies used to generate the big data and consequently their associated spatial and temporal characteristics (zeroth order properties) have an important effect on a set of metrics we are interested in for planning purposes (first and second order properties). Consequently, four important questions arise, and they are discussed in the remainder of this section.

As the underlying data generation process changes lead to changes in spatial and temporal properties, as well as changes in trip related metrics, how shall we interpret the resulting changes?

As noted earlier, the change from mobile phone data to app-based data resulted in a closer resemblance to household travel survey data for trip rates (3.23 from apps data vs 4.40 from the PSRC travel survey data, in comparison to 1.78 from mobile phone data vs 3.89 from the Buffalo travel survey data). And correlation with population density at the census tract level increased from 0.43 to 0.91. Clearly, improvement in data quality both in terms of locational accuracy (Figure 4) and temporal sparsity (Figure 12), resulting in more accurate calculation of metrics such as home census tracts and trip rates. But questions still remain: for frequency of observations, is better? Or is there a threshold after which the bias of under-estimation becomes that of over-estimation? In our discussion on the May 9th data shift (Section 3.4), we showed that from before to after May 9th, there was a 33 percent increase in the number of observations per ID (Figure 31) and consequently an improvement in temporal sparsity (Figure 34). The average trip rate consequently increased from 3.11 to 3.47, edging closer to the 4.4 from the household travel survey data. However, the difference was not apparent for activity duration (Figure 35), departure time (Figure 37), or trip length (Figure 38) distributions⁹. Given these observations, we provide a few responses to the above question. First, when temporal sparsity is relatively low (which is the case for both mobile phone data and app-based data) and locational accuracy is low (which is the case for mobile phone data), improvements in both will draw metrics closer to the ground truth, and this is in particular the case for locational accuracy. Second, however, as temporal sparsity continues to increase, the marginal benefit decreases. In fact, beyond a certain threshold, we expect the positive benefit can turn negative, although

⁹ The cumulative distribution for trip length (Figure 39), however, also indicated that the curve after May 9th was closer to that of the PSRC survey data.

this will require future research. Third, improvements in different metrics may vary, as shown in trip rate, activity duration, departure time, and trip length.

Reversely, one may also wonder about the likelihood that the improvement in trip rate estimation using data after May 9th (as compared to data before May 9th) was due to chance instead of increases in the number of observations and thus temporal sparsity. On the basis of what we learned from both the mobile phone data and the app-based data, it is extremely unlikely that the improvement observed in trip rate estimation was due to pure chance. From both data sets, we observed that first, under-estimation of trip rates is common in big data sets because of the temporal sparsity issue; and second, an initial increase in trip rate is present as temporal sparsity improves. It is worth noting that this assessment specifically applies to trip rate estimation, as the effects on other metrics are more complex. As shown in our analyses, increases in the number of observations appear to have little to no impacts on other metrics.

Can we be more proactive in estimating trip-related metrics as the technologies and other circumstances underlying the big data generation process change over time?

As evidenced from the sudden increase in the number of observations per ID after May 9th in the app-based data, the technologies used to generate the big data will inevitably change. Consequently, the associated data properties will change, as well as the estimated metrics we are interested in. While it is important to monitor how they may change over time, it is also an interesting question to ask whether we can be ahead of the changes by predicting the consequences of the changes, such as what we observed in the May 9th phenomenon. This question points to important future research directions that seek to establish linkages between zeroth order properties (data properties such as locational accuracy and temporal sparsity) and first- and second-order properties. Understanding the nature of these linkages will give us predictive capability.

How do we deal with the issue that big data lack ground truth?

As noted in Chen et al. (2014, 2016), because of the uncontrolled data generation process associated with big data, validation of the inferred statistics from the data is critically important. And yet, there is no ground truth data for most of the trip-related metrics. Therefore, frequently household travel surveys are used for validation purposes. Although they represent a very important first step in the right direction, note that the inferred results at the individual level can have large errors even though a high level of accuracy is observed at the aggregate level. The paper by Chen et al (Chen et al., 2016) discussed a number of ways to accomplish additional validation, including the use of simulation data (Chen et al., 2014), collection of small sample GPS/survey data, and using experiments and models to understand the effects of data properties (e.g., locational accuracy and temporal sparsity) on the metrics of interest (e.g., trip rate). Further investigations are critically needed for the validation of results generated by big data sources.

How can we make better data via integration of big and small data?

Besides being big, a very unique aspect of the big data is their continuous and dynamic nature, meaning that they are potentially available during any time and at any place. This is in stark

contrast to the small travel survey data that are static, capturing travel patterns on a typical day once every 5 to 10 years¹⁰. The static nature of the travel survey data renders them only useful for long-term (usually 20 to 30 years) demand forecasts but nearly useless in assessing many short-term and equally important policy and operation scenarios that arise frequently from time to time. As an example, understanding user profiles and their associated travel patterns in corridor management is critical not only for operation purposes (e.g., evaluating the effectiveness of tolling and other control strategies such as ramp metering) but also for policy evaluation and adjustment (e.g., understanding how different users and communities are affected by the control strategies provides basis for policy evaluation and adjustment). Big data, because of their dynamic and continuous nature, can be leveraged to provide answers to these important questions. This is the case especially when the big data are integrated with other data including, for example, household travel survey data, census data, flow data (e.g., travel volumes and speeds from loop detectors), and license plate data that are already collected by state or local DOTs. This data fusion exercise will not only result in better data that leverage the advantages of diverse data sets, but will also move us toward more real-time, continuous management of our transportation facilities based on the principles of efficiency, equity, and safety. The realization of this vision requires the development of data fusion frameworks and methodologies and their validation, which are currently nearly non-existent. In Section 5 of this report, a data fusion framework, including goal, objectives, and major considerations, is presented. Development and comprehensive testing of more specific, detailed data fusion methodologies are also highly recommended for future research.

¹⁰ Most travel surveys are conducted once every 10 years.

4.0 Other Emerging Data Sources and Applications

This section provides a summary of other data sources from emerging technologies and systems in transportation, and their potential applications. These include data from connected and automated vehicles (CAVs) and new shared mobility services. Notice that these emerging technologies, as well as the data they provide and the applications they support, are currently under rapid development. This section aims to provide a brief discussion of the technologies, the data they can provide, and the applications they can support. A more comprehensive and detailed survey of these technologies, their data, and applications are beyond the scope of this project.

4.1 Data from Connected Vehicles

Connected vehicles (CVs) are vehicles that can communicate (i.e., send and receive messages) with the surrounding environment, including other vehicles (defined as vehicle to vehicle (V2V) communication), infrastructures (defined as vehicle to infrastructure (V2I) communication), pedestrians (defined as vehicle to pedestrian (V2P) communication), and other entities (defined as vehicle to everything (V2X) communication). In 2011, the USDOT published the performance goals of the CV system (Campolo and Molinaro, 2013) based on the results from pilot deployment tests. The report showed that CV systems could save 1083 lives annually (Lee and Lim, 2012), and reduce up to 27 percent of time delays (Vinel, 2012) and 20 percent of gas emissions by deploying just two safety applications. These findings indicate that CV systems could be an effective solution for safety, mobility, and environmental problems in the current transportation system.

To enable CV, multiple communication technologies have been applied, such as dedicated short-range communications (DSRC), cellular networks (i.e., 3G/4G/5G), Wi-Fi, and radar. Figure 44 provides an illustration of the DSRC architecture, supported by a number of IEEE and SAE standards. The top layer, i.e., the Application Layer, concerns CV data and related applications. In particular, two SAE standards (J2735 and J2945.1) define the message sublayer of CV, which are the data items transmitted between vehicles and the surrounding environment (other vehicles, infrastructure, pedestrians, etc.). Safety and other applications can then be built on the CV data sets. For other communication technologies, the Application Layer, especially the message sublayer, remains the same. In the following, important CV message sets are introduced with some sample data provided. CV-related applications are to be discussed.

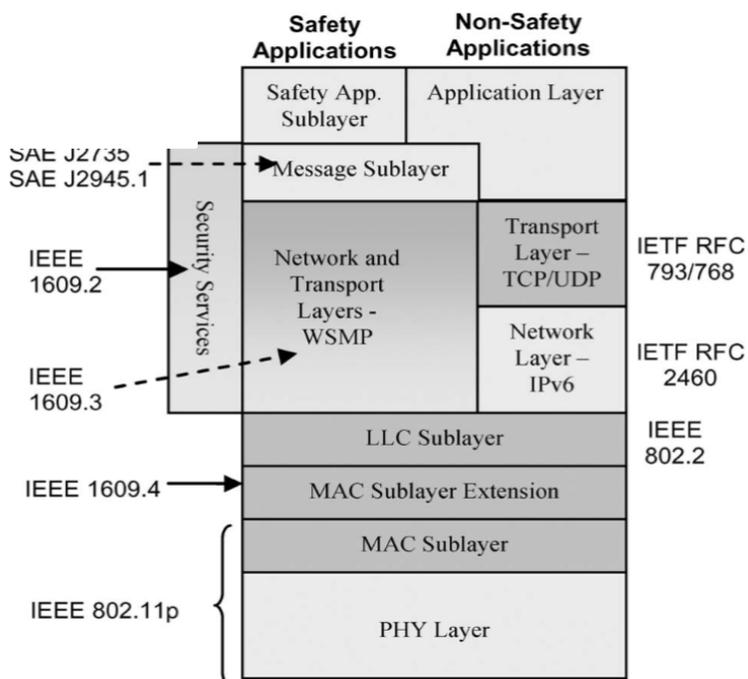


Figure 44 Graph. Layered architecture of dedicated short-range communications (DSRC) [12]

4.1.1 CV Data

Table 5 lists all the messages (15 messages in total) defined in SAE J2735_201603 (i.e., the latest version of J2735) (Dedicated Short Range Communications (DSRC) Message Set Dictionary, 2016). Every message shown in Table 5 can provide data for one major application area in the CV system. Additionally, the combination of messages can provide more information. For example, Basic Safety Message (BSM) and Map Data provide vehicle status information and road network information separately. However, the combination of BSM and Map Data can offer services such as left-turn assistance.

The table clearly shows that, in comparison to other “big data” from mobile devices (such as cellular data, GPS data, or app-based data), CV data provide much richer information about not only the location and speeds of individual vehicles (or other users such as pedestrians if V2X data are available), but also vehicle status (such as emissions), road maps, infrastructure data (such as signal timing), and hazards information, among others. As a result, a much wider range of applications can be supported by CV data, as presented in Section 4.1.2. More importantly, the data generation process of the CV data is much clearer and transparent to users than that of other big data sources, leading to a better understanding of the CV data. However, some of the important issues identified for big data (Section 3 of this report or the report by Chen et al. (2017)), such as representativeness of the underlying population, may also apply to CV data and thus need to be properly investigated and addressed. This will be especially true during the transition process of CV technology, when the penetration of CV-enabled vehicles will be low and which may take one or a few decades to complete (Lin, 2015).

In this subsection, to further show the detailed CV data set, two CV message sets, BSM and signal phase and timing (SPaT) data are presented. Some sample data are provided in the appendices.

Table 5. Messages defined in SAE J2735_201603 (Dedicated Short Range Communications (DSRC) Message Set Dictionary, 2016)

Message set		
Order #	Message set	Description
1	Basic Safety Message (BSM)	Exchange safety data regarding vehicle state
2	Map Data (MAP)	Convey many types of geo-road information
3	SPaT	Convey current status of one/more signalized intersections
4	Common Safety Request (CSR)	Provides a means: a veh participating in the exchange of BSM can unicast to other vehs for additional information
5	Emergency Vehicle Alert (EVA)	Broadcast emergency veh related warning message to surrounding vehs
6	Intersection Collision Avoidance (CICAS-V)	Limited to stop signs and traffic signal violations
7	NMEA corrections	Encapsulate NMEA 183 style differential corrections for GPS/GNSS radio navigation signals
8	Probe Data Management (PDM)	The type of data collected and sent by OBUs to the local RSU
9	Probe Vehicle Data (PVD)	Exchange status about a vehicle with other DSRC devices to allow the collection of information about typical vehicle traveling behaviors along a segment of road.
10	Road Side Alert (RSA)	Send alerts for nearby hazard to travelers
11	RTCM corrections	Encapsulate RTCM differential corrections for GPS and other radio navigation signals
12	Signal Status Message (SSM)	Relate the current status of the signal and the collection of pending/active preemption/priority requests acknowledged by the controller.
13	Traveler Information (TIM)	Send various types of information (advisory and road sign types) to equipped devices
14	Personal Safety Message (PSM)	Broadcast safety data regarding the kinematic state of various types of Vulnerable Road Users (VRU)
15	Test Message	Provide expandable messages for local and regional deployment use.

Source: www.sae.org

Basic Safety Message

BSMs are transmitted between infrastructure and vehicles at high frequency (typically 10 Hz), which allows safety critical applications such as collision avoidance. Table 6 (Araniti et al., 2013) shows the data format of the core BSM defined in SAE J2735 with brief descriptions of each data element (the extension BSM (Xu et al., 2017) is a supplement of the core BSM, which will not be discussed in the report). Appendix A.4.1 shows the sample data of core BSMs collected by the University of Michigan’s Transportation Research Institute (UMTRI) in 2012.

Table 6. Data frame of core Basic Safety Message (BSM) (Wyoming CV Pilot Basic Safety Message One Day Sample)

Field Name	Units	Description
FileID	None	Reference number to locate the source of the data in its original file
TxDevice	None	ID (number) of the device transmitting the BSM
Gentime	milliseconds	A more secure form of Epoch time, which is influenced by 1609.2 of the IEEE 1609 family of standards-related network management and security
TxRandom	None	Randomly assigned ID to mask the device ID of the transmitting device for security purposes
MsgCount	None	Message ID that gets incremented by one with each BSM
DSecond	Deciseconds	Time in deciseconds since ignition started
Latitude	Degrees	Current latitude of the vehicle
Longitude	Degrees	Current longitude of the vehicle
Elevation	Meters	Current elevation of the vehicle according to its GPS
Speed	m/sec	Vehicle speed
Heading	Degrees	Vehicle heading/direction
Ax	m/sec ²	Longitudinal acceleration
Ay	m/sec ²	Lateral acceleration
Az	m/sec ²	“Vertical” acceleration
Yawrate	Deg/sec	Vehicle yaw rate
PathCount	None	Number, between 1 and 23, representing a group of points that communicate a vehicle’s position and motion. Each group of points is of non-uniform size.
RadiusOfCurve	Centimeter	Estimation of the radius of a curve being negotiated, which is derived from a number of systems and sensors. Positive and negative values reflect right and left turns, respectively, and +/- 32767 for straight paths.
Confidence	Percent	Signals the accuracy and non-steady state and steady state of curvature estimate. In steady state (straight roadways or curves with the constant radius of curvature), a high confidence value is reported.

Source: data.transportation.gov

Signal Phase and Timing Data

A SPaT message is a bidirectional transmission message between infrastructure (traffic signals in this case) and vehicles. Traffic signals send this message to surrounding vehicles to inform them about the status of signal phasing and timing. Such information can help vehicles estimate travel times and select the most efficient routes. Meanwhile, vehicles can also send messages to nearby infrastructure (traffic signals in particular) to report their travel velocities and positions, which can assist a signal to detect traffic flow status within the neighborhood of the signal. As a result, signal systems can take actions on the basis of this given information to adjust their timing plan to improve the flow of traffic and reduce congestion. The data format of SPaT, defined in J2735_201603 (Dedicated Short Range Communications (DSRC) Message Set Dictionary, 2016), is shown in Table 7 (Araniti et al., 2013), with brief descriptions of each data element. Appendix A.4.2 provides some SPaT sample data.

Table 7. Data format of SPaT Messages

Field Name		Description
Name		Name of the intersection; to be used only in debugging
ID		A globally unique value set, consisting of a region ID and intersection ID assignment. Provides unique mapping to the intersection MAP in question, which provides complete location and approach/move/lane data
Revision		
Status		General status of the controllers
Moy		Minute of current UTC year, used only with messages to be archived
Timestamp		The mSec point in the current UTC minute that the message was constructed
EnabledLanes		A list of lanes where the Revocable bit has been set which are now active and therefore part of the current intersection
States		State name for the movements, to be used only in debugging
ManeuverAssistList	MovementName	
	SignalGroup	An index used to map the differences between the internal state machine of one or more signal controllers
	State-time-speed	Consisting of ets of movement data
	AssistList	Flow or traffic for the lanes and maneuvers in question
	ConnectionID	The common connectionID used by all lanes to which -- this data applies
	QueueLength	Unit = 1 meter, 0 = no queue
	WaitOnStop	If "true", the vehicles on this specific connecting -- maneuver have to stop on the stop-line and not enter the collision area
	PedBicycleDetct	True if ANY ped or bicycles are detected crossing -- the above lanes

Source: Data.gov

4.1.2 Applications

In the last decade, a few dozen CV-related applications have been developed. In particular, the USDOT sponsored the CV Pilot Deployment Program that grouped the applications into six categories: Safety Applications (both V2I safety and V2V safety), Agency Data, Environmental Applications, Dynamic Mobility Applications, Road Weather, and Smart Roadside (Intelligent Transportation Systems - CV Pilot Deployment Program), as shown in Table 8.

Table 8. The applications selected by the USDOT to utilize in a CV Pilot Program

V2I Safety	Environment	Mobility
Red Light Violation Warning (RLVW)	Eco-Approach and Departure at Signalized intersection	Advanced Traveler Information System
Curve Speed Warning	Eco-Traffic Signal Timing	Intelligent Traffic Signal System (I-SIG)
Stop Sign Gap Assist	Eco-Traffic Signal Priority	Signal Priority (transit, freight)
Spot Weather Impact Warning	Connected Eco-Driving	Mobile Accessible Pedestrian Signal System (PED_SIG)
Reduced Speed/Work Zone Warning	Wireless Inductive/Resonance Charging	Emergency Vehicle Preemption (PREEMPT)
Pedestrian in Signalized Crosswalk Warning(Transit)	Eco-Lanes Management	Dynamic Speed Harmonization (SPD-HARM)
	Eco-Speed Harmonization	Queue Warning (Q-WARN)
	Eco-cooperative Adaptive Cruise Control	Cooperative Adaptive Cruise Control (CACC)
	Eco-Traveler Information	Incident Scene Pre-Arrival Staging
	Eco-Ramp Metering	Guidance for Emergency Responders (RESP-STG)
	LOW Emissions Zone Management	Incident Scene Work Zone Alerts for Dryers and Workers (INC-ZONE)
	AFV Charging/Fueling Information	Emergency Communications and Evacuation (EVAC)
	ECO-Smart Parking	Connection Protection (T-CONNECT)
	Dynamic Eco-Routing (light vehicle, transit, freight)	Dynamic Transit operations (T-DISP)
	ECO-ICM Decision Support System	Dynamic Ridesharing (D-RIDE)
		Freight-Specific Dynamic Travel Planning and Performance
		Drayage Optimization
		Smart Roadside
		Wireless Inspection
		Smart Truck Parking
V2V Safety	Road Weather	
Emergency Electronic Brake Lights (EEBL)	Motorist Advisories and Warnings (MAW)	
Forward Collision Warning(FCW)	Enhanced MDSS	
Intersection Movement Assist	Vehicle Data Translator (VDT)	
Left Turn Assist(LTA)	Weather Response Traffic Information (WxTINFO)	
Blind Spot/Lane Change Warning (BSW/LCW)		
Do Not Pass Warning (DNPW)		
Vehicle Turning Right in Front of Bus Warning (Transit)		
Agency Data		
Probe-based Pavement Maintenance		
Probe-enabled Traffic Monitoring		
Vehicle Classification-based Traffic Studies		
CV-enabled Turning Movement & Intersection Analysis		
CV-enabled Origin-Destination Work Zone Traveler Information		

Source: www.its.dot.gov

Most of the applications in Table 8 have been tested in the real world (e.g., using CV testbeds) or in simulation studies, with their benefits and lessons learned summarized. The most noticeable study was probably the Safety Pilot Deployment Program by UMTRI, sponsored by the USDOT (Safety Pilot: Model Deployment). Both V2V and V2I applications were tested and demonstrated by the Safety Pilot Deployment Program. To illustrate, we briefly summarize two

V2V applications, including Intersection Movement Assistance (IMA) and Forward Collision Warning (FCW), and one V2I application on Red Light Violation Warning (RLVW). IMA is meant to warn a driver not to enter an intersection when high risks are detected by the sensors on his/her vehicle. For example, if a red light violation suddenly occurred at the intersection, the IMA feature would alert the driver of the danger via V2V communications. FCW is designed to detect risks and alert drivers to avoid possible collisions with front vehicles through appropriate actions. FCW warns drivers or takes automatic emergency actions when impending rear-end collisions are detected by the sensors from the rear vehicle. RLVW is a V2I application that enables a CV when it approaches a signalized intersection to receive information from the infrastructure regarding the geometry of the intersection and the signal timing. Along with the vehicle information of speed and acceleration, it is feasible to determine the likelihood with which the vehicle will run into a red light when it enters the intersection. If the violation seems highly likely to occur, then a warning can be provided to the driver (Red Light Violation Warning).

To help resolve the specific safety, mobility, and environment issues of agencies, the US DOT ITS Joint Program Office introduced the basic steps for the selection and implementation of CV application (CV102: Participant Workbook Sept 2015). There are three main steps in general to select and implement a CV application, as illustrated in the figure below, which are also briefly explained in this report.



Figure 45 Graph. Process to select applications

Source: CV102: Participant Workbook Sept 2015

Step 1: Identify Local Needs

This step addresses the problems and challenges in the local transportation system that an agency manages. It could range from extreme weather condition, emission concerns for certain areas in the city, to heavy congestion on a corridor, or intersection safety. For instance, a road section with sharp turns might require CV applications to help reduce the probability of car accidents.

Step 2: Set Performance Goals

After needs and issues have been identified, the purpose of step 2 is to set measurable goals for quantifying the target improvement the agency aims to achieve. Below are some examples of performance goals.

- Reduce crashes by 10 percent; injuries by 20 percent; and fatalities by 30 percent
- Reduce pedestrian-vehicle conflicts by 50 percent
- Ensure that transit vehicles are on schedule 90 percent of the time

- Increase peak period output by 8 percent
- Reduce emissions by 20 percent
- Reduce fuel costs associated with operating a transit fleet by 10 percent.

For specific agencies and issues, some or all of the above performance goals may apply with proper modifications, or additional performance goals may need to be developed.

Step 3: Select Applications

Step 3 is to select a specific CV application or a combination of CV applications from the list in Table 5 to meet the performance goals identified in Step 2 for solving the issues identified in Step 1. The selection process analyzes the issues each CV application aims to address, compares them with the local issues identified in Step 1, and identifies/compares the benefits of the selected applications to determine whether the performance goals identified in Step 2 can be satisfied. This process is expected to be interactive with Step 1 and Step 2, and iterative, with possible revisions and refinements to the local needs in Step 1 and the performance goals in Step 2 before the selection of the set of CV applications can be finalized.

4.2 Data from Automated Vehicles

Wikipedia defines automation as “the technology by which a process or procedure is performed without human assistance.” Automated vehicles (AV) are one subcategory of automated technology, as “self-governing” vehicles can navigate themselves through inputs of information collected from their surrounding environments without any human assistance (Xu et al., 2017). However, because of current technological limitations, scientists cannot completely automate vehicles. To officially measure the degree of automation in vehicles, SAE published J3016 to standardize the levels of driving automation and specify the definition of each level (“J3016_201806,”), as shown in Table 9. NHSTA also defines similar levels of automation (Automated Vehicles for Safety, 2017).

Table 9. Levels of driving automation (NHSTA & SAE)

Level (NHTSA/SAE)	Name (SAE)	Definition (NHTSA)	Definition (SAE)
0	No driving automation	Zero autonomy; the driver performs all driving tasks	The performance by the driver of the entire DDT, even when enhanced by active safety systems.
1	Driver assistance	Vehicle is controlled by the driver, but some driving assist features may be included in the vehicle design	The sustained and ODD-specific execution by a driving automation system of either the lateral or the longitudinal vehicle motion control subtask of the DDT (but not both simultaneously) with the expectation that the driver performs the remainder of the DDT
2	Partial driving automation	Vehicle has combined automated functions, like acceleration and steering, but the driver must remain engaged with the driving task and monitor the environment at all times	The sustained and ODD-specific execution by a driving automation system of both the lateral and longitudinal vehicle motion control subtasks of the DDT with the expectation that the driver completes the OEDR subtask and supervises the driving automation system.
3	Conditional driving automation	Driver is a necessity, but is not required to monitor the environment. The driver must be ready to take control of the vehicle at all times with notice.	The sustained and ODD-specific performance by an ADS of the entire DDT with the expectation that the DDT fallback-ready user is receptive to ADS-issued requests to intervene, as well as to DDT performance relevant system failures in other vehicle systems, and will respond appropriately.
4	High driving automation	The vehicle is capable of performing all driving functions under certain conditions. The driver may have the option to control the vehicle.	The sustained and ODD-specific performance by an ADS of the entire DDT and DDT fallback without any expectation that a user will respond to a request to intervene.
5	Full driving Automation	The vehicle is capable of performing all driving functions under all conditions. The driver may have the option to control the vehicle.	The sustained and unconditional (i.e., not ODD specific) performance by an ADS of the entire DDT and DDT fallback without any expectation that a user will respond to a request to intervene.

Source: www.sae.org

Presently, the technologies for AVs are under rapid development. Many industry leaders such as Google (Waymo), Uber, Tesla, GM, and Ford are working on AV technologies, testing, and user cases. They are collecting huge amounts of data, which are however rarely shared with researchers or the public. Furthermore, different from CV data, no data standards have been developed for AV data. A current project, “IEEE Standard: WG2040 - Standard for Connected, Automated and Intelligent Vehicles: Overview and Architecture Working Group” (Campolo and Molinaro, 2013), may touch on AV data standards issues. Before such standards are officially

released, one might expect that AV data may be similar to CV data as discussed above (possibly with additional data such as Lidar data and video data); if this is the case, then the AV-related applications and potential issues may also be similar to those of CV data. However, such expectations can only be proved true (or wrong) when official AV data standards and applications are released.

The only data set that is currently available for AVs is the AV-related accident database, which is released as requested by government agencies for safety reasons. For example, some AV accident data are publicly available as requested by Caltrans (California Department of Transportation) (Report of Traffic Collision Involving an Autonomous Vehicle (OL 316)). Table 10 shows the main data fields used when describing accidents involving AVs (Lee and Lim, 2012). A.4.3 in Appendix A.4 shows a sample of the accident data from one accident report.

Table 10. Autonomous vehicle accident data format [8]

Variables	Description
Time	Accurate time of the accident happened
Date	The date when the accident happened
Brand	The brand of the vehicle
Location	The location where the accident happened
Speed	The velocity of the vehicle when the accident happened
Accident Type	The type of accident
Police Called	Whether the vehicle called the police
Injured	Injuries from the accident
Responsibility	Was the accident caused by human interaction
State	The State to which the testbed belong
Note	Some notes for the accident

Source: Lee and Lim, 2012

4.3 Data from New Shared Mobility Services

This section summarizes some of the currently available data sources related to new shared mobility services, including data from ridesourcing and bike-sharing. The summary here does not mean to be exhaustive but provides some examples with formats and samples of the available data sets from these services.

First, the term of *shared mobility*, which can be traced back to the 1990s in North America, includes various forms of bike-sharing, carsharing, ridesharing (carpooling and vanpooling), and ridesourcing services. It is known as an innovative transportation strategy for users to have short-term access to transportation services as needed (Shared Mobility: Current Practices and Guiding Principles). The emerging forms of these services (i.e., *new shared mobility services*) are featured in app-based platforms, matching users and services to satisfy “on-demand” requests (e.g., the use of bikes or cars or ride services). In practice, there are several other terminologies defined for ridesourcing—the use of a platform to “source” rides from a driver pool (Shaheen et al., 2017)—such as transportation network company (TNC), e-hailing, ridehailing, e-booking, etc. To avoid confusion, in this report, the term “ridesourcing” is used, which is also recommended by the recent SAE standards on emerging shared mobility services (Shared and Digital Mobility Committee, 2018).

There is also some confusion regarding the differences between ridesourcing and ridesharing. Conceptually, ridesourcing is distinct from ridesharing. Ridesharing indicates that drivers are travelers who share similar origins/destinations with their riders for a common purpose of conserving resources, saving money, or saving time. Ridesourcing, on the other hand, is a for-hire commercial service and operates much like taxi services.

The use of app-based platforms help generate massive data sets related to new shared mobility services, including at least two types: those related to the requests of services (also called “order data”) and those related to the locations and movements of service vehicles or bikes (also called “trajectory data”). Because of the privacy concerns of users/drivers and the protection of their competitive advantages, new shared mobility providers have not been very enthusiastic about sharing their data. Therefore, available shared mobility data are quite limited at the current stage. In the following, we summarize the available data as those provided directly by the service providers (e.g., Uber, Lyft, Didi) and those via public data sharing platforms (e.g., Kaggle, GitHub).

4.3.1 Lyft Data

Instead of building a user interface for data sharing, Lyft established an application programming interface (API) for users so that they can request and receive Lyft data, including the operation area, geographic information (latitude and longitude), and the possible time durations for a selected trip (shown as Figure 46). To successfully operate Lyft API, users need to acquire a key by signing into the Lyft Developer platform and are also required to have basic programming capabilities (e.g., Java Script, Python).

To request data from Lyft API, users need to provide a csv file with needed information (e.g., pick-up locations, origins and destination) to the API and to develop a script to obtain the data they need. For instance, to request data for pick-up locations, including ride type, pick-up time estimate and nearby drivers, the input should be the latitude and longitude of the pick-up location.

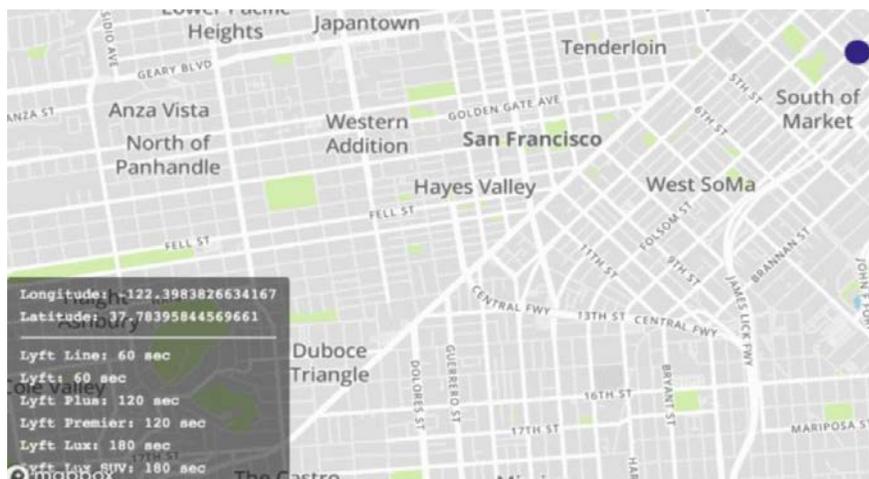


Figure 46. Lyft API

Source: <https://developer.lyft.com/>

4.3.2 Uber Data

In contrast, with consistent feedback from cities to use aggregated data for urban planners, Uber launched Uber Movement in 2017. Regarded as a planning tool, the initial goal of Uber Movement is to share historical traffic flow data (anonymized) for urban design to increase the efficiency of urban traffic. The data covered several cities across the world, including North America, Central and South America, Europe, Africa, Asia, Australia and New Zealand (e.g., Seattle, Bogota, London, Nairobi Mumbai, Sydney, Taipei, etc.). The user interface is similar to Figure 47.

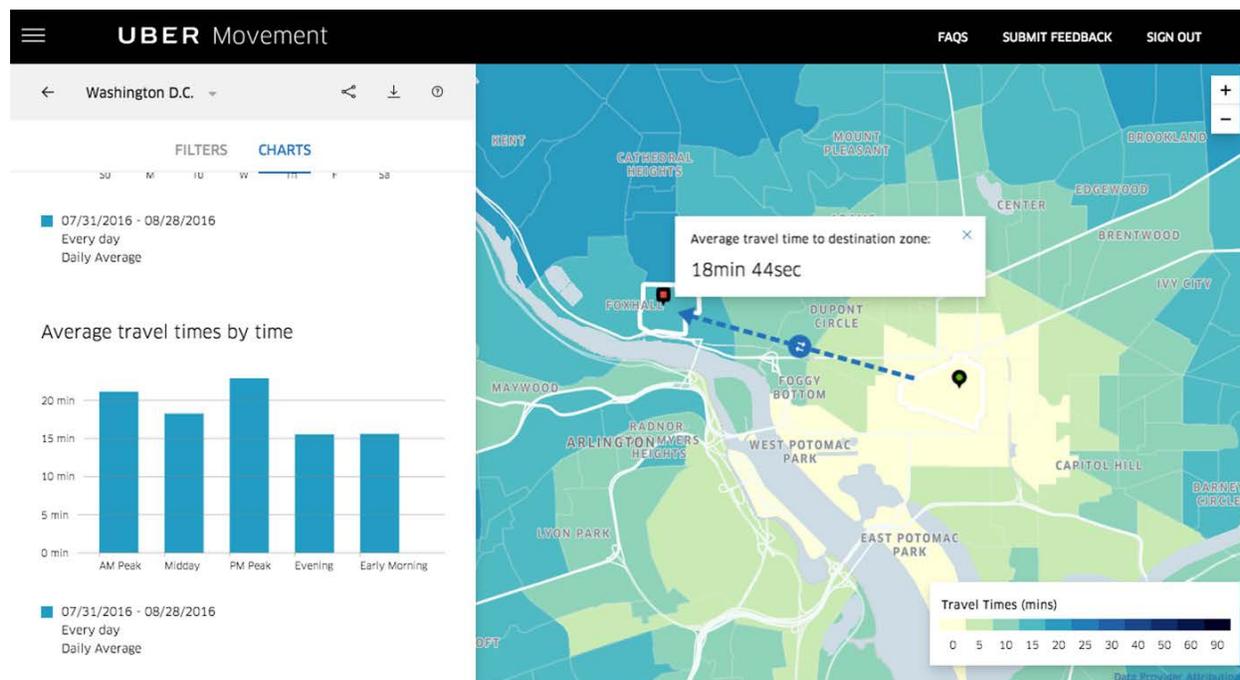


Figure 47. Uber Movement user interface

Source: <https://movement.uber.com/>

There are three main types of data sets in general from Uber Movement for users to download: FILTERED DATA, ALL DATA, and GEO BOUNDARIES (Uber Movement: Let’s find smarter ways forward).

To request the FILTERED DATA, as the name implies, a user needs to indicate where and when the data they would like to request by selecting the city, zone type (census tracts, traffic analysis zones), the date-time range (covering day of week: from Monday to Sunday, and time of day: daily average, AM peak, midday, PM peak, evening and early morning), and the origin and destination.

When the selection step has been completed, three data sets can be downloaded (sample data are shown in Appendix A.5.1):

- 1) Origin to all destinations: The data set includes the aggregated mean and range of travel time from the starting zone to all other zones.
- 2) Daily time series: Based upon the selected origin zone and destination zone, the data set covers means and ranges of travel times for: all day, AM peak, midday, PM peak, evening and early morning.
- 3) Chart data: On the basis of the selected origin zone and destination zone, the data set includes aggregated mean and range of travel times for 'day of week' and 'time of day'.

The ALL DATA category covers the arithmetic mean, geometric mean, and standard deviations for aggregated travel times over the selected data-range between each OD zone pair in the city. As of recently, the data can be downloaded from the first quarter in 2016 to the first quarter in 2018. The following data set files will be generated:

- Hourly Aggregate (all days)
- Hourly Aggregate (weekdays only)
- Hourly Aggregate (weekends only)
- Weekly Aggregate
- Monthly Aggregate (all days)
- Monthly Aggregate (weekdays only)
- Monthly Aggregate (weekends only).

The data formats for all the data files are very similar, as shown in Table 11 (Data Sample displayed in Appendix A.5.2).

Table 11. Data format

Field	Type	Description
sourceid	String	Origin zone ID
dsid	String	Destination zone ID
month/hod	Number	Month of a year/Hours of a day
mean_travel_time	Seconds	(Arithmetic) Mean travel time
standard_deviation_travel_time	Seconds	(Arithmetic) Standard deviation of travel time
geometric_mean_travel_time	Seconds	Geometric mean travel time
geometric_standard_deviation_of_travel_time	Seconds	Geometric standard deviation of travel time

Source: movement.uber.com

For GEO BOUNDARIES, a file with JavaScript Object Notation (JSON) format (JSON File), including geometric boundaries information for tract zones (with ID number addressed for each zone), is provided. The data can be viewed in GIS software packages (e.g., ArcGIS, QGIS).

It is clear that the data currently from Uber Movement are aggregated at the TAZ or census tract level to provide information of travel times and total demand, while individual order or trajectory data are not included. Such data would be useful for certain urban planning applications that do not require detailed individual information.

4.3.3 DiDi Data

DiDi Chuxing is a China-based ridesourcing service provider (Homepage - DiDi official website). The GAIA Initiative (The Gaia Initiative) is the data sharing platform for DiDi, which aims to advance transportation research and promote the application of scientific research, and to strengthen the ties among industry, government agencies, and university researchers. Unlike Uber data, which require no application for users to obtain access for data, DiDi data are only accessible for academic research and require an application for access.

Currently, four data sets are available on the GAIA website, which are from two major cities in China: Xi'an and Chengdu during the month October and November 2016. Generally, there are two types of data from the GAIA Initiative for each city: route (trajectory) data and ride request (order) data. The trajectory data format is shown in Table 12 (sample data are provided in Table 20 of **DiDi Data** in the appendices). The measurement interval of the track points is about 2 to 4 seconds.

Table 12. Data format (Trajectory Data)

Field	Type	Comment
Driver ID	String	Anonymized
Order ID	String	Anonymized
Time Stamp	String	Unix timestamp, in seconds
Longitude	String	GCJ-02 Coordinate System
Latitude	String	GCJ-02 Coordinate System

Note: The origin-destination data of the mentioned area are insignificant in comparison to the data of the whole city. In addition, they fail to reflect city-wide supply and demand.

Source: outreach.didichuxing.com

The format of Didi's order data is shown in Table 13. In comparison to the trajectory data, the order data are only accessible for the city of Chengdu for November 2016. The data cover the GPS information (latitude and longitude) of pick-up and drop-off locations, order IDs, and ride start/end times. The sample data can be viewed in Table 21 in the appendices.

Table 13. Data format (Order Data)

Field	Type	Comment
Order ID	String	Anonymized
Ride Start Time	String	Unix timestamp, in seconds
Ride Stop Time	String	Unix timestamp, in seconds
Pick-up Longitude	String	GCJ-02 Coordinate System
Pick-up Latitude	String	GCJ-02 Coordinate System
Drop-off Longitude	String	GCJ-02 Coordinate System
Drop-off Latitude	String	GCJ-02 Coordinate System

Source: outreach.didichuxing.com

4.3.4 Public Open Data Sets

Kaggle Data

Known as the ‘AirBnB’ for data scientists, Kaggle is the largest community of data scientists and machine learners around the world, offering a crowd-sourced platform for data training, challenge, and competition. Ridesourcing and bike-sharing are the two main categories of shared mobility data that can be found in Kaggle. There are also ridesourcing data for Lyft and Uber services; however, only origins (with timestamps and GPS information) for each trip are provided, which appears to be more limited than what Uber and Lyft provide directly and therefore are not discussed in this report.

Various data resources for bike-sharing can be found in Kaggle. This project selected bike-sharing data as an example to show the data format Kaggle provides. The data source selected here was the Capital Bikeshare program in Washington, D.C. The format of the chosen data set in Kaggle is shown below, with the sample data displayed in Appendix A.5.4.

Table 14. Data format (Bike Sharing Demand | Kaggle)

	Type	Comment
datetime	String	Hourly date + timestamp
season	Number	1 = spring, 2= summer, 3 = fall, 4 = winter
workingday	String	Whether the day is neither a weekend nor holiday
weather	String	1: clear, few clouds, partly cloudy 2: mist + cloudy, mist + broken clouds, mist + few clouds, mist 3: light snow, light rain + thunderstorm clouds, light rain + scattered clouds 4: heavy rain + ice pellets + thunderstorm + mist, snow + fog
temp	Number	Temperature in Celsius
atemp	Number	‘feels like’ temperature in Celsius
humidity	Number	Relative humidity
windspeed	Number	Wind speed
count	Number	Number of non-registered user rentals initiated
registered	Number	Number of registered user rentals initiated
count	Number	Number of total rentals

Source: www.kaggle.com

GitHub Data

GitHub (Intro of GitHub) is known as a website and cloud-based service assisting developers in storing and managing their code, as well as tracking and controlling changes, with version

control as a connected principle. GitHub has a summarized, currently available bike-sharing data set as well. As the format of the bike-sharing data provided in GitHub is very similar to that in Kaggle, details of the data format and sample data are not listed here. Bike-sharing data from eleven different countries around the world, including the U.S. and Australia, are available from GitHub (Bike sharing, 2018).

4.3.5 Applications

With the emergence and rapid development of new shared mobility services, numerous data have been generated, bringing tremendous potential to many areas, including transportation applications. This section briefly summarizes a few examples of how new shared mobility data can be applied to transportation applications.

First, the data from shared mobility can help transportation researchers and policy makers to better explore and understand urban travel/traffic patterns. The increasing availability of data in urban traffic networks will increase the possibility to examine traffic flow patterns on a large-scale roadway network, as well as to observe the evolution of regional travel patterns through data mining (Ma et al., 2015). For example, Alexander and González (2018) assessed the impact of ride-sharing on city-wide congestion using the mobile data by extracting average daily OD trips from mobile phone records and estimating the proportion of the non-auto and auto travelers among the trips. Altshuler et al (2017) proposed a dynamic travel network approach for modeling and estimating potential ridesharing utilization over time. They concluded that the significant volatility of the utilization of ridesharing over time indicated the reliability of estimating the impacts of ride-sharing with dynamic network analysis. Li et al. (2017) conducted a different-in-different analysis to explore the impact of Uber on urban congestion. There is no doubt that applying 'big data' from shared mobility to explore and understand larger-scale traffic network patterns (e.g., city-wide congestion patterns) has the potential to initiate a revolution in urban mobility planning.

Second, instead of using data to reveal traffic network flow and travel patterns, with diverse shared mobility appearing in daily transportation, some scholars have begun research to obtain a deeper understanding of travel and choice behavior in terms of ridesourcing, bikesharing, etc. Shaheen et al, compared the variance in usage patterns between ridesourcing and taxis and found that younger users were inclined to choose ride-sourcing. Shaheen et al. (2016) used survey data to examine the motivation and behavior of casual carpoolers in San Francisco to understand how user characteristics (e.g., demographic information, users' attitudes toward carpooling services) affected their choices in comparison to taxis. Zhang et al. (2018) applied 5-months of trip data from bike-sharing users in Zhongshan, China, to understand their travel behaviors. They identified that most bike trips are part of a trip chain of multiple trips. With a sound understanding of travel behavior in choosing shared mobility options, a more comprehensive view can be obtained for urban planners and designers to develop more efficient multimodal transportation network systems, including transit and new shared mobility modes.

Third, traffic signal control, which has mainly relied on manually collected data from traffic counts and/or sensor data from infrastructures (e.g., video cameras, loop detectors, radar

detectors), may have the ability to be revolutionized by the large amount of shared mobility data. For example, DiDi Chuxing, the largest ridesourcing service provider in China, has been working on ways to use transportation big data analytics and artificial intelligence (3 ways Didi's big data is improving China's traffic · TechNode, 2017) to solve global transportation and urban and environmental challenges. One of their focus areas is to optimize urban traffic signals by using ridesourcing data, especially the trajectory data from ridesourcing vehicles. Different from conventional detector data, trajectory data from ridesourcing vehicles serve as a low-cost, continuous, and reliable data source, which can help greatly improve conventional, detector-based signal control methods (Zheng et al., 2018). Over the last two years, their trajectory-based traffic signal control and optimization algorithms have been applied to hundreds of signalized intersections in a number of Chinese cities, leading to reduced congestion and improved travel times (Didi Chuxing CTO Keynotes Symposium). Such trajectory-based traffic signal timing optimization methods can also be co-developed with connected/automated vehicles (Li and Ban, 2018) to help build a more intelligent, efficient, and sustainable transportation system.

Fourth, apart from just addressing traffic mobility issues, shared mobility data can also be applied to improve roadway safety. For instance, the SIN (safety in numbers), mentioned by Jacobsen (2003), explored correlations among collision accidents with walkers and cyclists. Such research results can be combined with bikesharing and ridesharing data to explore their impacts on road safety. For example, Fishman and Schepers (2016) examined the influence of bikesharing programs on cycling safety with a combination of injury data, ridership data, and crash data. They concluded that bikesharing users are associated with fewer bicycle crashes (fatal/injury) than are private riding cyclists (using their own bikes). Morrison et al (2018) explored the correlation between ridesharing and motor vehicle crashes by using time-series analysis in four U.S. cities (Portland, Las Vegas, Reno, and San Antonio), considering time-sequential impacts from the usage of Uber and Lyft. They found that ridesharing may increase the total number of crashes; however, it may also reduce vehicle accidents due to drunk driving.

In summary, data from new shared mobility services will be helpful and useful for understanding the traffic/travel patterns of road networks and travel behaviors, and for improving traffic control and traffic safety. With more research efforts conducted using data from new shared mobility services, a more comprehensive view and understanding of the transportation system can be produced, which will help establish a more efficient, intelligent, and sustainable transportation ecosystem.

5.0 Development of A Data Fusion Framework

The analysis results of the app-based data in this report, as well as the results based on mobile phone data and GPS data in the team’s previous research (Chen et al., 2017), clearly demonstrate that big data, despite their great value for travel pattern analysis, do have their own issues, most noticeably uncertainty in the data generation process and related representativeness issues with respect to the underlying population. Indeed, both big data and small data have their unique characteristics, and therefore advantages and limitations, as summarized in Table 4 and briefly discussed in the last section. Table 15 presents more details about the characteristics of various types of big data and small data, further illustrating the values and pros/cons of different data sources.

More importantly, different datasets may complement each other. For example, small data (such as travel surveys) are often static (i.e., collected once a few years), whereas big data are mostly dynamic (able to be collected almost continuously); most big data show just traces of devices (or people), whereas small data often contain much richer information (such as the demographics of the underlying population). Therefore, it would be more beneficial to properly integrate big data and small data from different sources to create data with better quality (e.g., to alleviate bias issues). At the same time, different types of data may also have commonalities that can serve as the basis to link data sets for data fusion. Figure 48 shows the commonalities and differences among big data, small data (travel surveys in particular), and traditional flow data (e.g., from loop detectors). The figure shows the general relationships among different categories of data, while more specific commonalities and differences should be identified when actual data sets are encountered.

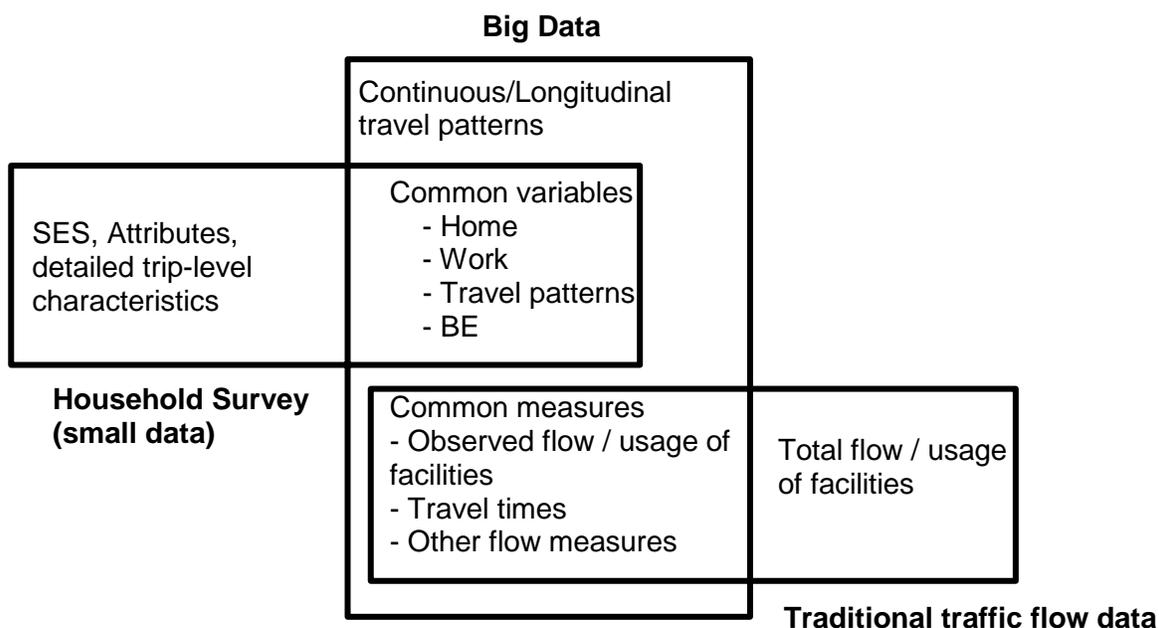


Figure 48 Graph. Integration of big data and small data

Table 15. Characteristics of different data sets

Datasets	Level	Variables				Modes	Properties		Types
		Locations	SES	Trip/ activity	BE		Locational accuracy	Temporal patterns	
Small data: HHS	Person-level	Yes	Yes	Yes	Yes	All	N/A	N/A	Cross-sectional/ panel
Big data: mobile phone	Person-level	Yes	No	No	No	All	Coarse	Sparse	Continuous
Big data: app-based	Person-level	Yes	No	No	No	All	Mixed (coarse and fine-grained)	Less sparse, clustered on roadways	Continuous
Big data: probe vehicle	Vehicle-level	Yes	No	No	No	Vehicle only	Fine-grained	Dense	Continuous
Big data: CAVs	Vehicle-level	Yes	No	No	No	Vehicle only	Fine-grained	Dense	Continuous
Traditional flow data	Aggr. level	Yes	No	No	No	All	Mixed	Dense	Continuous

In this section, a preliminary data fusion framework is proposed to combine big data and small data. The proposed framework is preliminary because it contains only the goal, objectives, basic principles, and important considerations for data fusion. Ways to develop more specific data fusion methods will require more in-depth investigation, which will be left for future research.

5.1 Goal and Objectives of Data Fusion

The main goal of data fusion is to produce better quality data and/or more complete data for given transportation planning or operational applications. Specific objectives of data fusion may be as follows:

- Improved data quality: Data from a single source may be subject to errors and/or biases, which can often be corrected or alleviated by merging data from multiple sources.
- Filling data gaps: Single-sourced data often have limited spatial coverage or observation periods (e.g., travel surveys conducted only for a few months), or the collected data are restricted for certain populations (e.g., vehicular GPS data are only for vehicles, while transit smart card data are only for transit users). Combining data from multiple sources can help provide data with more complete coverage (spatially, temporarily, or the user population).
- Validation: Data from different sources can help validate each other. This is particularly so when “ground truth” data are not available or are difficult to obtain. For example, travel surveys may be used to validate the trip-related analysis results from app-based data, as illustrated in Section 3.
- Analysis: Data fusion may help analysts better understand and interpret analysis results (given a lack of ground-truth data).

5.2 Principles of Developing Data Fusion Methods

There are a few considerations in developing data fusion methods. First, data fusion method should be developed on the basis of the target application. There is rarely a pure data fusion method without considering any application, since the purpose of data fusion is to provide better data for certain application. This application-centric view is important since in some cases, it is possible that a simple combination of the datasets from different sources may be sufficient (e.g., by providing data for different aspects of the problem), in which case a rigorous data fusion method (as we propose here) may not be necessary.

Second, according to the specific application, key *performance measures* should be defined to assess the performance/success of the application. For example, travel time and reliability may be used for a tolling project of a key urban corridor, while the percentage of single occupancy vehicle (SOV) travelers may be used for a project of adding a new transit line.

Third, proper *data quality metrics* should also be defined to help quantify the quality of each data source. The metrics may be defined to measure the data quality in terms of accuracy, timeliness, spatial/temporal coverage, representativeness, etc. Depending on the objectives and performance measures of the specific application, different data quality metrics may need to be

defined for the same data source. The application, performance measures, and data quality metrics can together establish certain data standards that can help assess the data needs of the given application and its associated data requirement.

The fourth step is to analyze the properties of each data source (especially emerging big data sources) to have an in-depth understanding of the properties of the data, their pros and cons, and the quality of the data based on the defined data quality metrics. For example, for regional travel pattern analysis using big data, one may apply the analysis framework in Section 3 to analyze the properties of each big data source. This step is critical to develop proper data fusion methods that can leverage the advantages of all data sources, while at the same time controlling their limitations to acceptable levels.

Furthermore, the following aspects may also be useful when data fusion methods are developed:

- Understanding use profiles, such as where people live and work and where they come and go, can be extremely important for some applications.
- To address bias issues, synthetic data may be helpful (Rodriguez et al., 2018), and in certain cases, more rigorous bias modeling and correction methods (Zagheni and Weber, 2015) need to be developed to address data biases more effectively
- Validation of the data fusion methods is important. For this, ground truth data (or benchmark data if ground truth data are not available) are crucial for validation purposes. Here benchmark data can be understood as data sources that are known to have relatively higher quality in certain aspect (e.g., location accuracy, representativeness, etc.).
- The field of transportation has been extensively studied in the past, resulting in well-established theories and models (collectively referred to as “transportation knowledge” in this report). Data, big or small, are not expected to fundamentally change most of such knowledge; rather they should reflect the knowledge (or help reveal new knowledge in certain cases). Data fusion methods therefore should adequately consider both data and proper knowledge of the application. On the one hand, to deal with massive, often heterogeneous data sources, data-driven methods are crucial, including machine learning algorithms and especially recent deep learning based methods. On the other hand, to respect established knowledge in transportation, suitable models may also need to be integrated with data driven methods. Such data-driven, model-based methods can be important alternatives for developing transportation data fusion methods in the future.
- Ultimately, data fusion is just one way to address data related issues and to improve data quality. It is certainly not the only way, and in some cases, may not even be the best way. Therefore, **understanding the data and application is of paramount importance to whether and how to develop data fusion methods.**

6.0 References

- 3 ways Didi's big data is improving China's traffic · TechNode, 2017. TechNode.
- Alexander, L.P., González, M.C., Assessing the Impact of Real-time Ridesharing on Urban Traffic using Mobile Phone Data 9.
- Altshuler, T., Katoshevski, R., Shifan, Y., 2017. Ride Sharing and Dynamic Networks Analysis. ArXiv170600581 Phys.
- Araniti, G., Campolo, C., Condoluci, M., Iera, A., Molinaro, A., 2013. LTE for vehicular networking: a survey. *IEEE Commun. Mag.* 51, 148–157. <https://doi.org/10.1109/MCOM.2013.6515060>
- Automated Vehicles for Safety [WWW Document], 2017. NHTSA. URL <https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety> (accessed 11.13.18).
- Bike sharing, 2018. . BetaNYC.
- Bike Sharing Demand | Kaggle [WWW Document], URL <https://www.kaggle.com/c/bike-sharing-demand/data> (accessed 10.10.18).
- Calabrese, F., Lorenzo, G.D., Liu, L., Ratti, C., 2011. Estimating Origin-Destination Flows Using Mobile Phone Location Data. *IEEE Pervasive Comput.* 10, 36–44. <https://doi.org/10.1109/MPRV.2011.41>
- Campolo, C., Molinaro, A., 2013. Multichannel communications in vehicular Ad Hoc networks: a survey. *IEEE Commun. Mag.* 51, 158–169. <https://doi.org/10.1109/MCOM.2013.6515061>
- Chen, C., Ban, X. (Jeff), Wang, F., Wang, J., Siddique, N., Fan, R., Lee, J., 2017. Understanding GPS and Mobile Phone Data for Origin-Destination Analysis. FHWA.
- Chen, C., Bian, L., Ma, J., 2014. From traces to trajectories: How well can we guess activity locations from mobile phone traces? *Transp. Res. Part C Emerg. Technol.* 46, 326–337. <https://doi.org/10.1016/j.trc.2014.07.001>
- 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. *Transp. Res. Part Policy Pract.* 44, 830–840. <https://doi.org/10.1016/j.tra.2010.08.004>
- 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. *Transp. Res. Part C Emerg. Technol.* 68, 285–299. <https://doi.org/10.1016/j.trc.2016.04.005>
- Cottrill, C.D., Pereira, F.C., Zhao, F., Dias, I.F., Lim, H.B., Ben-Akiva, M.E., Zegras, P.C., 2013. Future Mobility Survey: Experience in Developing a Smartphone-Based Travel Survey in Singapore. *Transp. Res. Rec.* 2354, 59–67. <https://doi.org/10.3141/2354-07>
- CV102: Participant Workbook Sept2015, Data.gov [WWW Document], Data.gov. URL <https://www.data.gov/> (accessed 11.27.18).
- Dedicated Short Range Communications (DSRC) Message Set Dictionary, 201603.
- Didi Chuxing CTO Keynotes Symposium | [WWW Document], URL <http://www.umtri.umich.edu/what-were-doing/news/didi-chuxing-cto-keynotes-symposium> (accessed 11.25.18).
- Fan, Y., Chen, Q., Liao, C.-F., Douma, F., 2013. UbiActive: A Smartphone-Based Tool for Trip Detection and Travel-Related Physical Activity Assessment 19.
- Fishman, E., Schepers, P., 2016. Global bike share: What the data tells us about road safety. *J. Safety Res.* 56, 41–45. <https://doi.org/10.1016/j.jsr.2015.11.007>
- Homepage - DiDi official website [WWW Document], URL <https://www.didiglobal.com/> (accessed 11.24.18).
- Intelligent Transportation Systems - CV Pilot Deployment Program [WWW Document], URL https://www.its.dot.gov/pilots/cv_pilot_apps.htm (accessed 10.10.18).

- Intro of GitHub [WWW Document], . Kinsta Manag. WordPress Hosting. URL <https://kinsta.com/knowledgebase/what-is-github/> (accessed 10.22.18).
- Iqbal, M.S., Choudhury, C.F., Wang, P., González, M.C., 2014. Development of origin–destination matrices using mobile phone call data. *Transp. Res. Part C Emerg. Technol.* 40, 63–74. <https://doi.org/10.1016/j.trc.2014.01.002>
- J3016_201806,
- Jacobsen, P.L., 2003. Safety in numbers: more walkers and bicyclists, safer walking and bicycling. *Inj. Prev.* 9, 205–209. <https://doi.org/10.1136/ip.9.3.205>
- JSON File [WWW Document], URL <https://fileinfo.com/extension/json> (accessed 11.13.18).
- Kaggle [WWW Document], URL <https://www.kaggle.com/> (accessed 10.21.18).
- Lee, S., Lim, A., 2012. Reliability and performance of IEEE 802.11n for vehicle networks with multiple nodes, in: 2012 International Conference on Computing, Networking and Communications (ICNC). Presented at the 2012 International Conference on Computing, Networking and Communications (ICNC), pp. 252–256. <https://doi.org/10.1109/ICCNC.2012.6167422>
- Li, W., Ban, X.J., 2017. Traffic signal timing optimization in connected vehicles environment, in: 2017 IEEE Intelligent Vehicles Symposium (IV). Presented at the 2017 IEEE Intelligent Vehicles Symposium (IV), pp. 1330–1335. <https://doi.org/10.1109/IVS.2017.7995896>
- Li, Z., Hong, Y., Zhang, Z., An empirical analysis of on-demand ride-sharing and traffic congestion 10.
- Liao, C.-F., Chen, C., Fan, Y., 2017. A Review on the State-of-the-Art Smartphone Apps for Travel Data Collection and Energy Efficient Strategies. Presented at the Transportation Research Board 96th Annual Meeting Transportation Research Board.
- Lin, L., 201508. Platoon Identification System in Connected Vehicle Environment 105.
- Ma, X., Yu, H., Wang, Yunpeng, Wang, Yinhai, 2015. Large-Scale Transportation Network Congestion Evolution Prediction Using Deep Learning Theory. *PLOS ONE* 10, e0119044. <https://doi.org/10.1371/journal.pone.0119044>
- Michalowski, T., 2017 Puget Sound Regional Travel Study 81.
- Morrison, C.N., Jacoby, S.F., Dong, B., Delgado, M.K., Wiebe, D.J., 2018. Ridesharing and Motor Vehicle Crashes in 4 US Cities: An Interrupted Time-Series Analysis. *Am. J. Epidemiol.* 187, 224–232. <https://doi.org/10.1093/aje/kwx233>
- Red Light Violation Warning [WWW Document], URL <https://local.iteris.com/cvria/html/applications/app57.html> (accessed 11.13.18).
- Report of Traffic Collision Involving an Autonomous Vehicle (OL 316) [WWW Document], URL https://www.dmv.ca.gov/portal/dmv/detail/vr/autonomous/autonomousveh_ol316+ (accessed 11.13.18).
- Rodriguez, L., Salimi, B., Ping, H., Stoyanovich, J., Howe, B., 2018. MobilityMirror: Bias-Adjusted Transportation Datasets. *arXiv:1808.07151 [cs]*.
- Safety Pilot: Model Deployment [WWW Document], URL <http://safetypilot.umtri.umich.edu/index.php?content=about> (accessed 11.13.18).
- Schewel, L., 2017. Location-Based Services Data Beats Cellular on Spatial Precision. *StreetLight Data*.
- Shaheen et,al, App-Based, On-demand Ride Services: Comparing Taxi and Ridesourcing Trips and User Characteristics in San Francisco.
- Shaheen, S.A., Chan, , Gaynor, T., 2016. Casual carpooling in the San Francisco Bay Area: Understanding user characteristics, behaviors, and motivations. *Transp. Policy* 51, 165–173. <https://doi.org/10.1016/j.tranpol.2016.01.003>
- Shared and Digital Mobility Committee, J3163_201809. SAE International. https://doi.org/10.4271/J3163_201809

- Shared Mobility: Current Practices and Guiding Principles [WWW Document], URL <https://ops.fhwa.dot.gov/publications/fhwahop16022/index.htm> (accessed 10.10.18).
- Stopher, P.R., 1996. Household travel surveys: new concepts and research needs, in: Conference on Household Travel Surveys: New Concepts and Research Needs. Conference Proceedings.
- Stopher, P.R., Greaves, S.P., 2007. Household travel surveys: Where are we going? *Transp. Res. Part Policy Pract.* 41, 367–381. <https://doi.org/10.1016/j.tra.2006.09.005>
- The Gaia Initiative [WWW Document], URL <https://outreach.didichuxing.com/research/opendata/en/> (accessed 10.10.18).
- Uber Movement: Let's find smarter ways forward [WWW Document], URL <https://movement.uber.com/?lang=en-US> (accessed 10.10.18).
- Vinel, A., 2012. 3GPP LTE Versus IEEE 802.11p/WAVE: Which Technology is Able to Support Cooperative Vehicular Safety Applications? *IEEE Wirel. Commun. Lett.* 1, 125–128. <https://doi.org/10.1109/WCL.2012.022012.120073>
- Wang, J., Wang, F., Cynthia, C., Ban, J.X., 2019. Comparative analysis of big and small data for deriving human mobility patterns. Presented at the Transportation Research Board 98th Annual Meeting Transportation Research Board, Appear soon.
- Wolf, J., Oliveira, M., Thompson, M., 2003. Impact of underreporting on mileage and travel time estimates: Results from global positioning system-enhanced household travel survey. *Transp. Res. Rec. J. Transp. Res. Board* 189–198.
- Wyoming CV Pilot Basic Safety Message One Day Sample - Data.gov [WWW Document], URL <https://catalog.data.gov/dataset/wyoming-cv-pilot-basic-safety-message-one-day-sample> (accessed 11.27.18).
- Xu, Z., Li, X., Zhao, X., Zhang, M.H., Wang, Z., 2017. DSRC versus 4G-LTE for Connected Vehicle Applications: A Study on Field Experiments of Vehicular Communication Performance [WWW Document]. *J. Adv. Transp.* <https://doi.org/10.1155/2017/2750452>
- Zagheni, E., Weber, I., 2015. Demographic research with non-representative internet data. *International Journal of Manpower.* <https://doi.org/10.1108/IJM-12-2014-0261>
- Zhang, Y., Brussel, M.J.G., Thomas, T., van Maarseveen, M.F.A.M., 2018. Mining bike-sharing travel behavior data: An investigation into trip chains and transition activities. *Comput. Environ. Urban Syst.* 69, 39–50. <https://doi.org/10.1016/j.compenvurbysys.2017.12.004>
- Zheng, J., Sun, W., Huang, S., Shen, S., Yu, C., Zhu, J., Liu, B., Liu, H.X., 2018. Traffic Signal Optimization Using Crowdsourced Vehicle Trajectory Data. Presented at the Transportation Research Board 97th Annual Meeting Transportation Research Board.

Appendixes

A.1 Appendix A—Extracting Trips from the App-based Data

We develop a ‘Divide, Conquer and Integrate’ (DCI) framework to extract trips from app-based data. In this appendix, we describe three steps of the DCI framework to extract trips from the app-based data: (1) Partition the data into data sets each of which contains smaller variance in spatiotemporal properties; (2) Extract trips from each data set independently by applying methods in accordance with the characteristics of the data set; (3) Combine trips extracted from all data sets by designing and applying a novel algorithm.

A.1.1 Partition data into low-variance sets

A stay is usually identified if the device does not move (e.g., more than 5 meters for GPS data and 1000 meters for cellular data) in a certain amount of time (e.g., 5 minutes). However, the variations embedded in a multi-sourced data suggest the definition of stays shall be variable as well: given the bimodal distribution of location accuracy in the app-based data, there exists no universal spatial constraint to define a stay. This can be illustrated in Figure 49, where one individual visited three places (l_0 , l_1 and l_2). Observations recorded at l_2 have worse location accuracy than those at l_0 and l_1 and therefore appear dispersed in space. To identify the stay at l_2 , one may want to define a stay as the device does not move farther than R_1 within, for example, 5 minutes. However, if the same definition is applied to l_0 and l_1 , the two stays could be mistaken as one when $R_1 > 2R_0$.

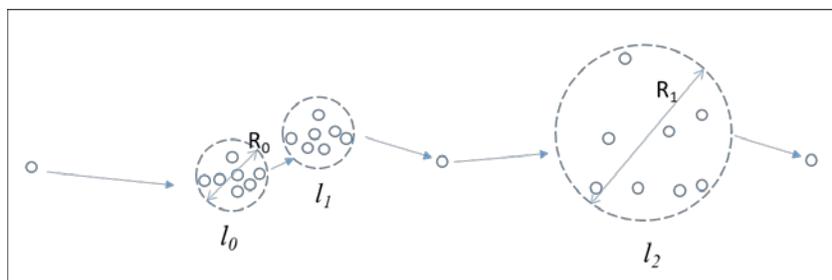


Figure 49. Graph. Illustration of variable definitions for stay identification

The data is partitioned such that each partition has small variance, based on which methods can be developed and applied to each data set. We observe that the app-based data is dominated by observations of high location accuracy, which appears similar to the GPS data and could be handled by a GPS-data-based method. Therefore, we partition the app-based data into two sets: one set contains observations with location accuracy no worse than a threshold p and the other set takes the remaining observations. In our study, we use $p=100$ meters which follows the distribution of location accuracy.

A.1.2 Extract trips from each data set independently

The partition step results in two data sets. The first one is similar to GPS data that can be processed using GPS-data-based methods. The second one contains observations with

location accuracy distributing around 1000 meters, which resembles cellular data. This observation is consistent with the data generation process, where cellular towers were used to locate a device when other technologies were not available. We thus propose to address the second data set using methods that are developed for cellular data. For clarity, hereafter, we refer to the first data set as the “GPS data set” and the second one as the “cellular data set”. In the following sections, the two data sets are processed independently to extract trips.

Extract trips from the GPS set

A commonly-used trace-segmentation method (Hariharan and Toyama, 2004; Ye et al., 2009) is applied to extract stays from the GPS set. For each trajectory of one user $\{d_1(t_1; lng_1, lat_1), d_2(t_2; lng_2, lat_2), \dots, d_i(t_i; lng_i, lat_i)\}$ ($t_1 \leq t_2 < \dots \leq t_i$), we extract stays by scanning through the trajectory and segmenting it into multiple sequences of observations with two parameters: signal roaming distance Δl_{roam} and the stay duration Δt_{dur} . A stay is extracted as a sequence of observations $\{d_m(t_m; lng_m, lat_m), d_{m+1}(t_{m+1}; lng_{m+1}, lat_{m+1}), \dots, d_n(t_n; lng_n, lat_n)\}$ ($t_1 \leq t_m \leq t_{m+1} < \dots < t_n \leq t_i$) satisfying both parameters: the distance between any two observations in the sequence should be shorter than Δl_{roam} and the duration (i.e. the time difference between the last and the first observation of this sequence $t_n - t_m$) must be no less than Δt_{dur} . This can be achieved by following the algorithm proposed in (Hariharan and Toyama, 2004). In our study, we use 200 meters and five minutes as Δl_{roam} and Δt_{dur} , respectively. This five-minute threshold follows the rule used in many household travel surveys to define what counts an activity (Transportation Research Board, 2005) and is used as an appropriate threshold for an activity location in the activity based modeling context (Yin et al., 2017). Δl_{roam} is set as 200 meters such that a displacement of 200 meters in five minutes corresponds to half of average walking speed 0.7 m/s, which is commonly used to distinguish between a stay and a movement (Bernardin, 2017). We replace locations in the sequence with their centroid (lng_c, lat_c) . Then, a sequence of observations representing a stay is simplified as $s_c(t_m, t_{m+1}, \dots, t_n; lng_c, lat_c)$.

We notice that stays representing multiple visits at a single place (e.g., one building) at different time are essentially unique in the form of longitude and latitude coordinates (Figure 50b). This prevents analyzing travelers’ mobility patterns such as regular returns to certain places (e.g., home, workplaces). Therefore, after the stay identification, we find those *common stays* that represent multiple visits to a single place. We achieve this by ignoring the temporal scale of stays and aggregating those close in space via an agglomerative clustering algorithm (Jiang et al., 2013). Specifically, we put together all stays identified in one user’s trajectories, aggregate those close in space into one cluster and replace locations of those stays (i.e. their centroids) with the centroid of the cluster (Figure 50c). Then, a stay $s_c(t_m, t_{m+1}, \dots, t_n; lng_c, lat_c)$ is modified as $s_{cc}(t_m, t_{m+1}, \dots, t_n; lng_{cc}, lat_{cc}; r_{cc})$, where the location (lng_{cc}, lat_{cc}) is the centroid of the cluster where s_c belongs. And r_{cc} records the radii of the cluster (the longest distance from the centroid to any stays in the cluster) as the locational uncertainty of s_{cc} . The data structure $s_{cc}(t_m, t_{m+1}, \dots, t_n; lng_{cc}, lat_{cc}; r_{cc})$ will be useful in the last step of our DCI framework. In the study, we apply the agglomerative clustering starting from each stay as individual cluster and set 200 meters (the same as the previous definition of roaming distance Δl_{roam}) as the criterion to stop the algorithm. Figure 50 illustrates the process identifying stays from the GPS data set, where one

common stay visited on two days, which is identified by aggregating two stays that are found in two trajectories.

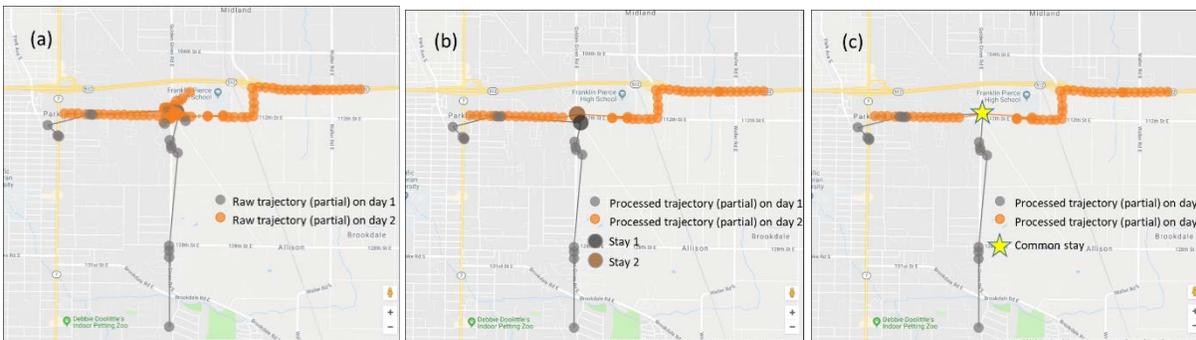


Figure 50. Illustration. Illustration of identifying stays from the GPS data set. (a) Raw GPS trajectories of two days; (b) Processed trajectories with identified stays; (c) Processed trajectories with a common stay being identified.

Source: Google Map

Extract trips from the cellular set

Given the low location accuracy and sampling frequency of the cellular data set, the trace-segmentation method that is designed for GPS data is not suitable. A framework developed by Wang and Chen (2018) for cellular data is applied in the study. The framework addresses the locational uncertainty and temporal sparsity of the cellular data with a revised incremental clustering method, which takes advantage of the longitudinal nature of the data. Following the method, we put together all cellular observations belonging to one user as a list \mathbf{d} and the list is clustered without regarding their time ordering:

- 1) starting from an observation d_0 , one new cluster C_0 is created and d_0 is the center;
- 2) each observation that has not been clustered will be checked and the one within a distance R_c to the center of C_0 is clustered into C_0 and the center of C_0 is correspondingly updated;
- 3) if no observation could be aggregated in the current cluster, one new cluster is created containing a non-clustered observation.

This procedure repeats itself until all observations in \mathbf{d} are clustered. This clustering returns a set of clusters, each of which contains observations that are close in space. Then, we come back to the time-ordered trajectories where temporal information is used such that a stay is identified as a sequence of observations within the same cluster and with duration exceeds a given threshold T_c (set as five minutes following the one for GPS data). Similarly, a stay is represented by $s_k(t_i, t_{i+1}, \dots, t_j; lng_c, lat_c; r_c)$, where (lng_c, lat_c) and r_c are the centroid and the radii of the cluster containing the sequence of observations, respectively. Through aggregating observations that are close in space but may be far away in time (e.g., several days), this method is able to identify common stays visited on multiple days. The spatial constraint R_c in

the algorithm is set as 1000 meters following the data characteristic observed in Figure 4, which is also used in previous studies on cellular data (Wang and Chen, 2018; Widhalm et al., 2015).

A.1.3 Integrating trips extracted from all data sets

We design an algorithm to integrate trips from the two data sets by referring to concepts of space-time relationships analyses in Geographic Information System (GIS) (Longley et al., 2005). Each data set is treated as a layer and identified stays as features in the layer. The time and location information (i.e. centroid and radius) of each stay act as the temporal and spatial attributes, respectively. Then, features (i.e., stays) from multiple layers (i.e., data sets) are combined by measuring their spatiotemporal relationship based on their temporal and spatial attributes.

For the app-based data, we use the predominant GPS data set as the basis. Then, for each of the cellular stay, we check its relationship with the processed GPS trajectory (observations of the same user on the same day), and decide how to combine it into the GPS trajectory. In the following, we define the temporal and spatial relationship, respectively.

The temporal relationship is defined in three categories:

- 1) Temporally separate: given a cellular stay $a(t_{a1}, t_{a2}, \dots, t_{ai}; lng_a, lat_a; r_a)$ and time-ordered GPS stays $\{\dots, b(t_{b1}, t_{b2}, \dots, t_{bj}; lng_b, lat_b; r_b), c(t_{c1}, t_{c2}, \dots, t_{ck}; lng_c, lat_c; r_c), \dots\}$ that are neighbors of a in time, we say a is temporally separate with GPS stays if $t_{a1} > t_{bj}$ and $t_{ai} < t_{c1}$ (Figure 51a).
- 2) Temporally contained: given a cellular stay $a(t_{a1}, t_{a2}, \dots, t_{ai}; lng_a, lat_a; r_a)$, if there exists a GPS stay $b(t_{b1}, t_{b2}, \dots, t_{bj}; lng_b, lat_b; r_b)$ such that $t_{b1} < t_{a1} < t_{ai} < t_{bj}$, we say a is temporally contained in b (Figure 51b).
- 3) Temporally intersected: given a cellular stay $a(t_{a1}, t_{a2}, \dots, t_{ai}; lng_a, lat_a; r_a)$, if it satisfies neither 1) nor 2), we say a is temporally intersected with GPS stays (Figure 51c).

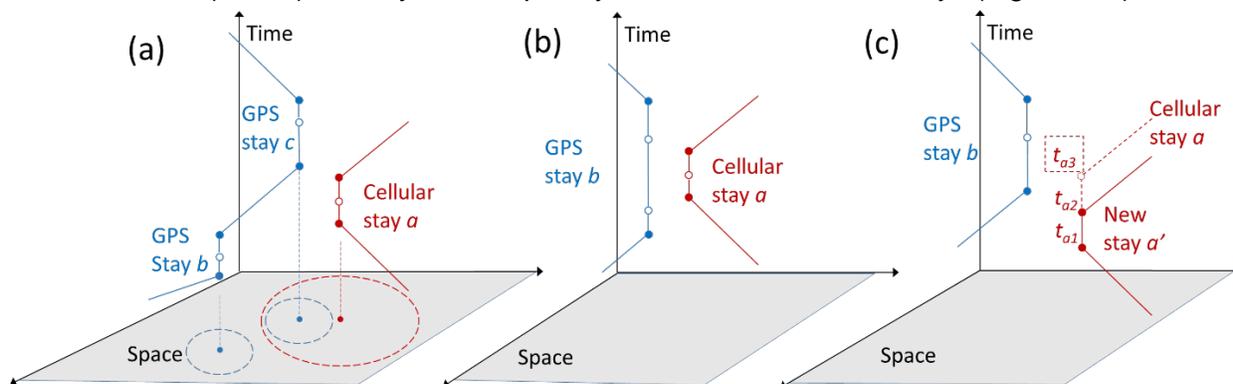


Figure 51. Illustration. Demonstration of the spatiotemporal relationship. (a) Temporally separate and spatially contiguous; (b) Temporally contained (c) Temporally intersected (cutting off t_{a3} turns a into a' which is temporally separate with b).

For the spatial relationship, we check whether the cellular stay is spatially contiguous with GPS stays or not. Here, two stays are defined spatially contiguous if the difference of their location uncertainty is greater than their spatial distance. Figure 52 gives an example where the stay a is

spatially contiguous with b , as the difference between their uncertainty radius (i.e. $r_a - r_b$) is greater than the distance between centroids of the two stays D_{ab} .

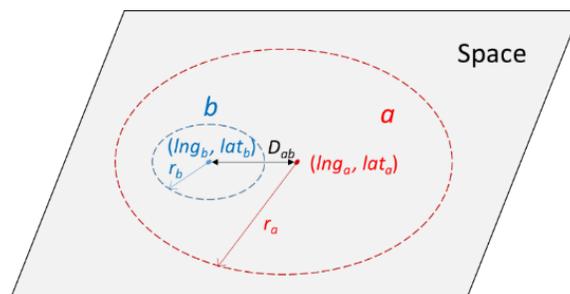


Figure 52. Illustration. Definition of spatially contiguous.

Following these definitions, the spatiotemporal relationship of each cellular stay with GPS stays is decided, based on which decision of the integration is made. For a cellular stay $a(t_{a1}, t_{a2}, \dots, t_{ai}; lng_a, lat_a; r_a)$:

- i) If it is temporally separate with its neighboring GPS stays b and c , a would either be a visit to a new place, or be the same visit at b or c but with a coarser location representation. This depends on whether a is spatially contiguous with b or c . If not, a is added as a new stay; otherwise, a is combined with b or c by replacing the location of a with that of b or c , depending on which one a is spatially contiguous with. Figure 51a gives an example where a is temporally separate and spatially contiguous with b .
- ii) If it is temporally contained in one GPS stay b , similarly, we check whether it is spatially contiguous with b or not. If yes, a is discarded; otherwise, a is inserted as a new stay and b could be split into two stays.
- iii) If it is temporally intersected with one (or more) GPS stay b , the intersected time period of a is cut off (Figure 51c), resulting into either a temporal separate or contained case. Then procedure i) or ii) is followed. The underlying logic of the cutoff is that, for the intersected time period, we use location information in b rather than in a , as locations in b have better accuracy and are more reliable. As illustrated in Figure 51c, the cutoff modifies a to a' , which is temporally separate with b . Therefore, we follow procedure i) to combine the new cellular stay a' into GPS trajectory.

The relation-checking and integration process repeats itself until all cellular stays are processed. Since the duration of some stays would change during the integration, we scan through each combined trajectory to update the duration of stays.

A.2 Appendix B—OD Estimation Method for App-Based Data

The zone-level observed trips from app-based data are usually not a good representation of the actual trips by the entire population due to at least two reasons: 1) the app-based dataset was not probabilistically-sampled so that it would not represent the pattern of the entire population well; and 2) because of the passively-solicited data generation process and data sparsity, app-based data may not capture all travels generated by the population; in other words, missing trips always exist for such datasets.

Admitting such issues in the dataset, one could still conduct a preliminary exploration on the OD estimation from the app-based data using the zone-level population data as a critical input. The steps of the OD estimation process based on the app-based data are briefly described below.

1) Aggregate residents into the TAZ level

Residents refer to the users whose home census tracts can be identified from app-based data due to frequent visits during night times. For the entire period, the total residents for each TAZ will be counted and added it as a new attribute associated with TAZs.

2) Calculate the scaling factors associated with each TAZ

The residents identified from the app-based data are samples from the entire population of each TAZ. The scaling factors can be calculated by the following equation:

$$\alpha_i = \frac{P_i}{r_i}$$

Where α_i denotes the scaling factor of TAZ i , and P_i , r_i correspond to the population and number of residents of TAZ i , respectively. All residents associated with TAZ i own the same scaling factor, equaling to α_i .

3) Generate OD matrix

Select the weekday trips generated by residents from the entire trip file. Multiply the trip with corresponding scaling factors and then assign it into OD matrix.

$$OD_{(a,b)} = \sum_i Trip_{(i,a,b)} * \alpha_i$$

Where OD is a matrix with the dimension of 3,700*3,700 (3700 is the total number of TAZs in the Puget Sound Region); a and b denote the trip origin and destination TAZ; and $Trip_{(i,a,b)}$ denotes the number of observed trips between TAZ pair (a, b) as well as generated by the user associated with TAZ i . Divide the OD matrix by the total number of weekdays involved, the daily OD demand matrix can be derived.

A.3 Appendix C—Home Distribution of Anonymous Users Observed Every Day

In the Section 3, we show that some IDs have every day observed and some IDs have a long life span. In this appendix, we provide more information on these IDs. Figure 53 gives a spatial distribution of home census tracts of IDs observed everyday (8,758 IDs). The distribution is compared with the population from the census. The comparison yields a correlation coefficient of 0.93 (Figure 54).

Similarly, Figure 55 gives a spatial distribution) of home census tracts of IDs with life span of 63 days (41,640 users). The distribution is also compared with the population from the census. The correlation coefficient is 0.98 (Figure 56).

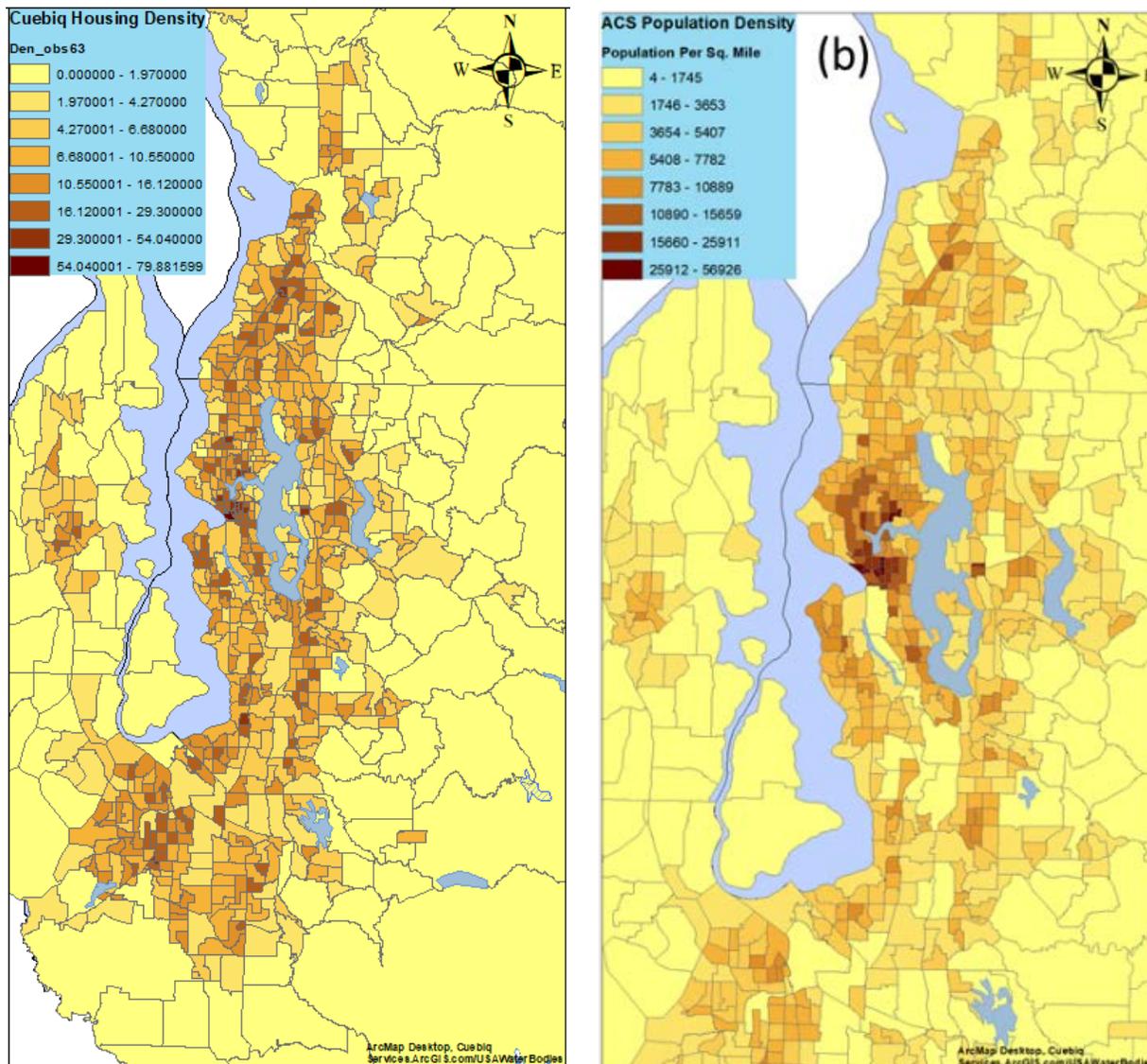


Figure 53. Map. Comparison between home census tracts of IDs observed everyday (8,758 IDs) and the population from the census. (a) Home density of IDs observed every day and (b) Census population density.

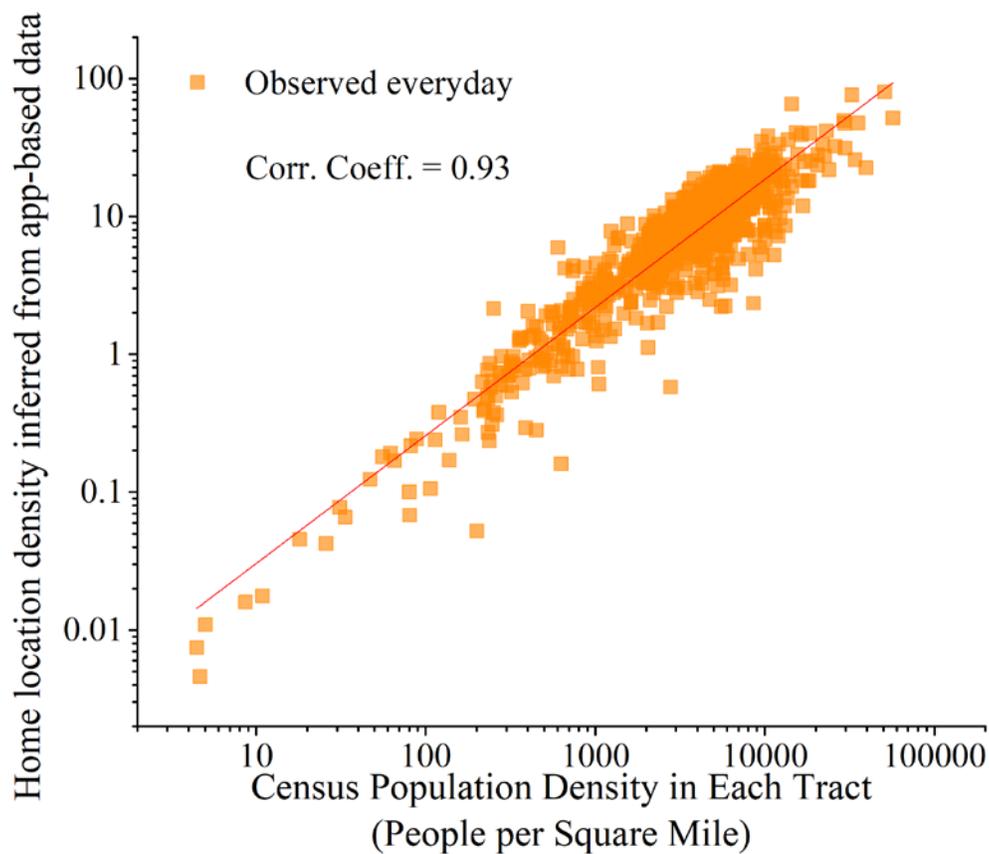


Figure 54. Graph. Correlation between home census tracts of IDs observed every day and census population (both at census tract level).

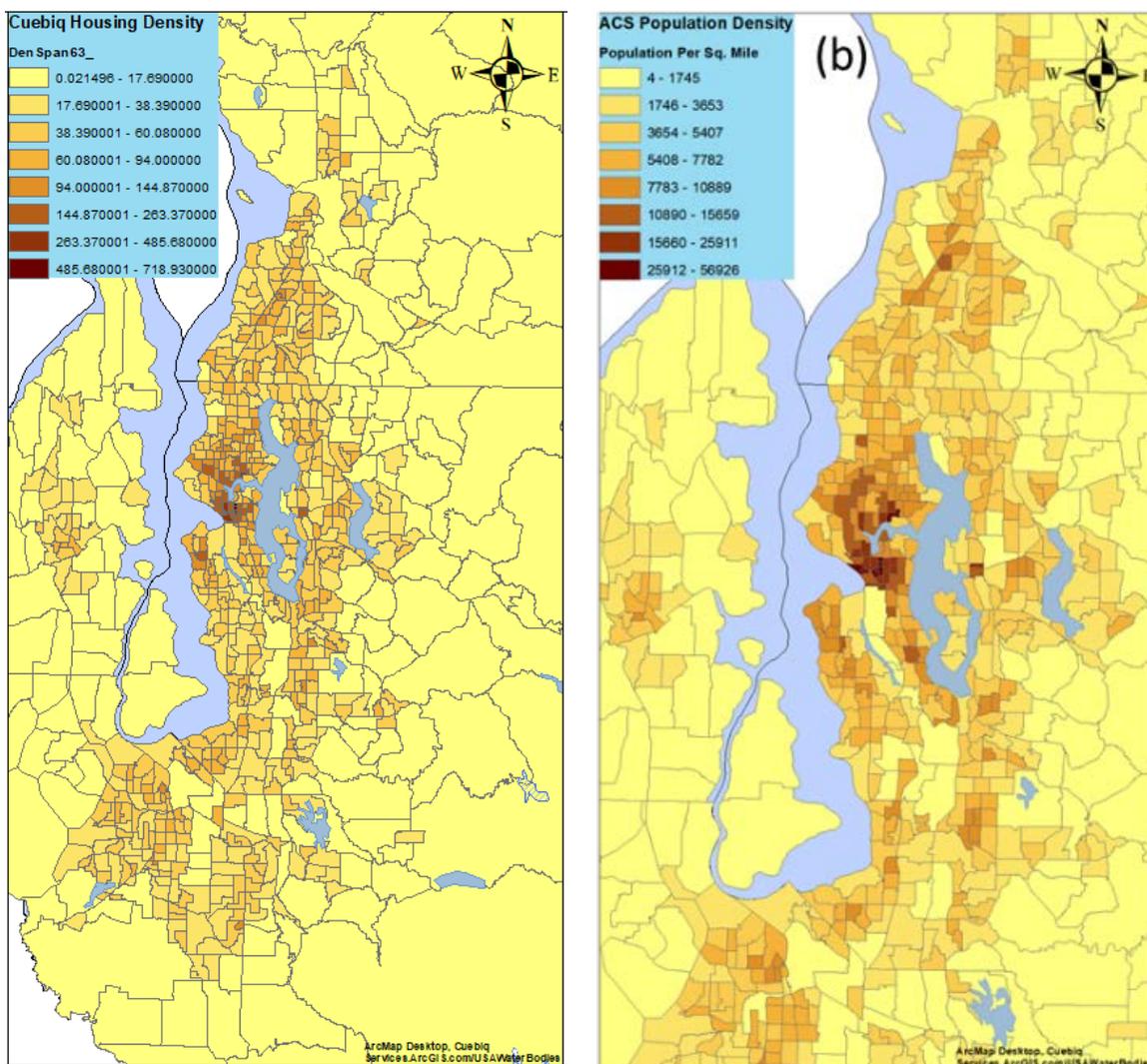


Figure 55. Map. Comparison between home census tracts of anonymous users with life span of 63 days (41,640 IDs) and the population from the census. (a) Home density of IDs with life span of 63 days presented at census tract level and (b) Census population density.

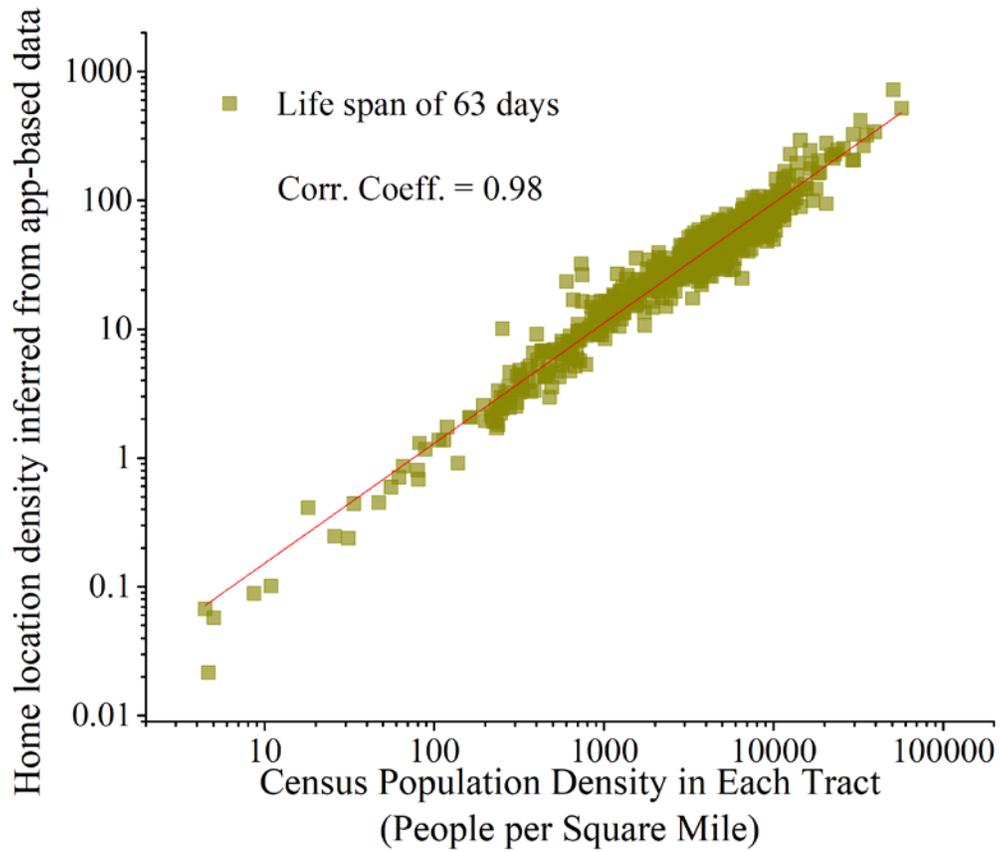


Figure 56. Comparison of home census tracts of IDs with life span of 63 days (41,640 IDs) at census tract level

A.4 Appendix D—CV data

A.4.1 Sample Data of BSM

FileId	TxDevice	Gentime	TxRandom	MsgCount	DSecond
13963	10	278,802,340,808,876.00	0	76	14700
13963	10	278,802,340,908,860.00	0	77	14800
13963	10	278,802,341,008,885.00	0	78	14900
13963	10	278,802,341,108,882.00	0	79	15000
13963	10	278,802,341,208,958.00	0	80	15100
13963	10	278,802,341,309,002.00	0	81	15200
13963	10	278,802,341,408,935.00	0	82	15300
13963	10	278,802,341,508,966.00	0	83	15400
13963	10	278,802,341,608,941.00	0	84	15500
13963	10	278,802,341,708,937.00	0	85	15600

Latitude	Longitude	Elevation	Speed	Heading	Ax	Ay	Az
42.29717	-83.7013	239.4	0.86	9.9375	-1.07	0.01	-10
42.29717	-83.7013	239.4	0.72	9.9375	-1.15	0.01	-10
42.29717	-83.7013	239.4	0.66	9.9375	-1.07	0.01	-10
42.29717	-83.7013	239.4	0.52	9.9375	-1.07	0.01	-10
42.29717	-83.7013	239.4	0.46	9.9375	-0.91	0.01	-10
42.29718	-83.7013	239.3	0.38	9.9375	-0.68	0.01	-10
42.29718	-83.7013	239.3	0.3	9.9375	-0.52	0.01	-10
42.29718	-83.7013	239.3	0	9.9375	-0.45	0.01	-10
42.29718	-83.7013	239.3	0.14	9.9375	-0.29	0.01	-10
42.29718	-83.7013	239.3	0	9.9375	-0.45	0.01	-10

Yawrate	PathCount	RadiusOfCurve	Confidence
-0.6	11	3276.7	100
-1.1	11	3276.7	100
-0.5	11	3276.7	100
-0.5	11	3276.7	100
-0.69	11	3276.7	100
-0.3	11	3276.7	100
-0.1	11	3276.7	100
-0.3	11	3276.7	100
0	11	3276.7	100
-0.1	11	3276.7	100

Source: data.transportation.gov

A.4.2 Sample Data of SPAT

MovementId	SPATID	Current State	Min Timeremaining	Max Timeremaining
3680586804	3040841724	0x04	362	1018
3680586824	3040841724	0x04	67	293
3680586845	3040841724	0x04	208	948
3680586848	3040841732	0x04	361	1017
3680586869	3040841724	0x01	147	643
3680586872	3040841732	0x04	66	292
3680586890	3040841724	0x40	208	704
3680586894	3040841732	0x04	207	947
3680586917	3040841724	0x40	656	1201
3680586918	3040841732	0x01	146	642

YellowState	YellowTime	Pedestrian Detect	Vehicle Pedestrian Count	LaneSet
NULL	0	0	0	0x01010504
NULL	0	0	0	0x02010701
NULL	0	0	0	0x03010904
NULL	0	0	0	0x01010504
0x02	36	0	0	0x04010B01
NULL	0	0	0	0x02010701
NULL	0	0	0	0x06020A02
NULL	0	0	0	0x03010904
NULL	0	0	0	0x0802
0x02	36	0	0	0x04010B01

Source: Data.gov

A.4.3 Sample data of AV accident

Time	Date	Brand	Location	Speed	Type	Police Called	Injured
21:27	1/8/18	GM	Intersection	0-20 mph	Side-impact collision	No	No
9:34	12/7/17	GM	Road Section	0-20mph	Sideswipe collisions	Yes	Yes
22:05	11/13/17	GM	Intersection	0-20 mph	Sideswipe collisions	No	No
AM	8/11/17	Navya	Road Section	0-20 mph	Sideswipe collisions	No	No
21:09	10/26/17	GM	Road Section	0-20 mph	Rear-end collision	No	No
9:16	10/20/17	GM	Intersection	NA	Sideswipe collisions	No	No
16:06	10/18/17	GM	Intersection	0-20 mph	Rear-end collision	No	No

Responsibility	State	Note
No	CA	With pedestrian
No	CA	With motorcycle
No	CA	
No	NE	
No	CA	
No	CA	Go through the intersection
No	CA	

Source: Lee and Lim, 2012



DMV
DEPARTMENT OF MOTOR VEHICLES
A Public Service Agency

REPORT OF TRAFFIC COLLISION INVOLVING AN AUTONOMOUS VEHICLE

DMV USE ONLY

AVT NUMBER

NAME

Instructions: Please print within the spaces and boxes on this form. If you need to provide additional information on a separate piece of paper(s) or you include a copy of any law enforcement agency report, please check the box to indicate "Additional Information Attached."

- Write **unk** (for unknown) or **none** in any space or box when you do not have the information on the other party involved.
- Give insurance information that is complete and which correctly and *fully* identifies the **company** that issued the insurance policy or surety bond, or whether there is a certificate of self-insurance.
- Place the National Association of Insurance Commissioners (NAIC) number for your Insurance or Surety Company in the boxes provided. The NAIC number should be located on the proof of insurance provided by you company or you can contact your insurer for that information.
- Identify any person involved in the accident (driver, passenger, bicyclist, pedestrian, etc) that you saw was injured or complained of bodily injury or know to be deceased.
- Record in the PROPERTY DAMAGE line any damage to telephone poles, fences, street signs, guard post, trees, livestock, dogs, buildings, parked vehicles, etc., including a description of the damage.
- Once you have completed this report, please mail to: Department of Motor Vehicles, Occupational Licensing Branch, P.O. Box 932342, MS: L224, Sacramento, CA 94232-3420

SECTION 1 — MANUFACTURER'S INFORMATION

MANUFACTURER'S NAME Waymo LLC	AVT NUMBER
BUSINESS NAME Waymo LLC	TELEPHONE NUMBER ()
STREET ADDRESS	CITY STATE ZIP CODE

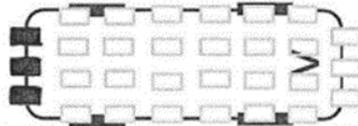
SECTION 2 — ACCIDENT INFORMATION/VEHICLE 1

DATE OF ACCIDENT 04/06/2018	TIME OF ACCIDENT 12:17 <input type="checkbox"/> AM <input checked="" type="checkbox"/> PM	VEHICLE YEAR 2017	MAKE Chrysler	MODEL Pacifica
LICENSE PLATE NUMBER	VEHICLE IDENTIFICATION NUMBER	STATE VEHICLE IS REGISTERED IN		
ADDRESS/LOCATION OF ACCIDENT Grant Road and Covington Road	CITY Mountain View	COUNTY Santa Clara	STATE CA	ZIP CODE 94040
Vehicle was: <input type="checkbox"/> Moving <input checked="" type="checkbox"/> Stopped in Traffic	Involvement in the Accident: <input type="checkbox"/> Pedestrian <input type="checkbox"/> Bicyclist <input type="checkbox"/> Other	NUMBER OF VEHICLES INVOLVED 2		
DRIVER'S FULL NAME (FIRST, MIDDLE, LAST)	DRIVER LICENSE NUMBER	STATE	DATE OF BIRTH	
INSURANCE COMPANY NAME OR SURETY COMPANY AT TIME OF ACCIDENT	POLICY NUMBER			
COMPANY NAIC NUMBER	POLICY PERIOD FROM TO			

Describe Vehicle Damage

UNK NONE MINOR
 MOD MAJOR

Shade in Damaged Area



DL 316 (REV. 2/2017) WWW



Figure 57. Sample accident report of AV (1)

SECTION 3 — OTHER PARTY'S INFORMATION/VEHICLE 2			
VEHICLE YEAR 0	MODEL Mercedes ML 350		
LICENSE PLATE NUMBER	VEHICLE IDENTIFICATION NUMBER unknown	STATE VEHICLE IS REGISTERED IN	
Vehicle was: <input checked="" type="checkbox"/> Moving <input type="checkbox"/> Stopped in Traffic	Involved in the Accident: <input type="checkbox"/> Pedestrian <input type="checkbox"/> Bicyclist <input type="checkbox"/> Other _____	NUMBER OF VEHICLES INVOLVED 2	
DRIVER'S FULL NAME (FIRST, MIDDLE, LAST) Unknown - driver left scene - reported to the Police		DRIVER LICENSE NUMBER unk	STATE 00
INSURANCE COMPANY NAME OR SURETY COMPANY AT TIME OF ACCIDENT unk		DATE OF BIRTH UNK	
COMPANY NAIC NUMBER unk	POLICY NUMBER unk		
	POLICY PERIOD FROM UNK TO UNK		
<input type="checkbox"/> Additional information attached.			
SECTION 4 — INJURY/DEATH, PROPERTY DAMAGE			
NAME (FIRST, MIDDLE, LAST)			
ADDRESS		CITY	STATE ZIP CODE
CHECK ALL THAT APPLY <input type="checkbox"/> Injured <input type="checkbox"/> Deceased <input type="checkbox"/> Driver <input type="checkbox"/> Passenger <input type="checkbox"/> Bicyclist <input type="checkbox"/> Property			
NAME (FIRST, MIDDLE, LAST)			
ADDRESS		CITY	STATE ZIP CODE
CHECK ALL THAT APPLY <input type="checkbox"/> Injured <input type="checkbox"/> Deceased <input type="checkbox"/> Driver <input type="checkbox"/> Passenger <input type="checkbox"/> Bicyclist <input type="checkbox"/> Property			
PROPERTY DAMAGE			
PROPERTY OWNER'S NAME		TELEPHONE NUMBER ()	
STREET ADDRESS		CITY	STATE ZIP CODE
WITNESS NAME		TELEPHONE NUMBER ()	
STREET ADDRESS		CITY	STATE ZIP CODE
WITNESS NAME		TELEPHONE NUMBER ()	
STREET ADDRESS		CITY	STATE ZIP CODE
<input type="checkbox"/> Additional information attached.			
SECTION 5 — ACCIDENT DETAILS - DESCRIPTION			
<input checked="" type="checkbox"/> Autonomous Mode <input type="checkbox"/> Conventional Mode			
A Waymo autonomous vehicle ("Waymo AV") in autonomous mode was rear-ended while stopped at a red light at the intersection of north Grant Rd and Covington Rd in Mountain View, CA. The Waymo AV was stopped for approximately 9 seconds when a vehicle approaching from behind made contact with the rear bumper of the AV at approximately 3 mph. There were no injuries reported at the scene by either party. The police were notified that the driver of the other vehicle left the scene without exchanging vehicle and insurance information.			
<input type="checkbox"/> Additional information attached.			

OL 316 (REV. 2/2017) WWW

Figure 58. Sample accident report of AV (2)

ITEMS MARKED BELOW FOLLOWED BY AN ASTERISK (*) SHOULD BE EXPLAINED IN THE NARRATIVE							
WEATHER (MARK 1 to 2 ITEMS)	VEH 1	VEH 2	MOVEMENT PRECEDING COLLISION	VEH 1	VEH 2	OTHER ASSOCIATED FACTOR(s) (MARK ALL APPLICABLE)	
A. CLEAR	✓	✓	A. STOPPED	✓		A. CVC SECTIONS VIOLATED CITED <input type="checkbox"/> YES <input type="checkbox"/> NO	
B. CLOUDY			B. PROCEEDING STRAIGHT		✓		
C. RAINING			C. RAN OFF ROAD				
D. SNOWING			D. MAKING RIGHT TURN				
E. FOG/VISIBILITY			E. MAKING LEFT TURN				
F. OTHER			F. MAKING U TURN			B. VISION OBSCUREMENT	<input type="checkbox"/>
G. WIND			G. BACKING			C. INATTENTION*	<input type="checkbox"/>
LIGHTING			H. SLOWING/STOPPING			D. STOP & GO TRAFFIC	<input type="checkbox"/>
A. DAYLIGHT	✓	✓	I. PASSING OTHER VEHICLE			E. ENTERING/LEAVING RAMP	<input type="checkbox"/>
B. DUSK - DAWN			J. CHANGING LANES			F. PREVIOUS COLLISION	<input type="checkbox"/>
C. DARK - STREET LIGHTS			K. PARKING MANUEVER			G. UNFAMILIAR WITH ROAD	<input type="checkbox"/>
D. DARK - NO STREET LIGHTS			L. ENTERING TRAFFIC			H. DEFECTIVE WEH EQUIP	CITED <input type="checkbox"/> YES <input type="checkbox"/> NO
E. DARK - STREET LIGHTS NOT FUNCTIONING*			M. OTHER UNSAFE TURNING				
ROADWAY SURFACE			N. XING INTO OPPOSING LANE			I. UNINVOLVED VEHICLE	<input type="checkbox"/>
A. DRY	✓	✓	O. PARKED			J. OTHER*	<input type="checkbox"/>
B. WET			P. MERGING			K. NONE APPARENT	<input type="checkbox"/>
C. SNOWY - ICY			Q. TRAVELING WRONG WAY			L. RUNAWAY VEHICLE	<input type="checkbox"/>
D. SLIPPERY (MUDDY, OILY, ETC.)			R. OTHER*				
ROADWAY CONDITIONS (MARK 1 TO 2 ITEMS)			TYPE OF COLLISION				
A. HOLES, DEEP RUT*			A. HEAD-ON				
B. LOOSE MATERIAL ON ROADWAY			B. SIDE SWIPE				
C. OBSTRUCTION ON ROADWAY*			C. REAR END	✓			
D. CONSTRUCTION - REPAIR ZONE			D. BROADSIDE				
E. REDUCED ROADWAY WIDTH			E. HIT OBJECT				
F. FLOODED*			F. OVERTURNED				
G. OTHER*			G. VEHICLE/PEDESTRIAN				
H. NO UNUSUAL CONDITIONS	✓	✓	H. OTHER*				

SECTION 6 — CERTIFICATION

I certify (or declare) under penalty of perjury under the laws of the State of California that the foregoing is true and correct.

I further certify that I am the authorized Administrator of the program for the above named employer.

PROGRAM DIRECTOR/AUTHORIZED REPRESENTATIVE PRINTED NAME AND TITLE MATTHEW SALWASSER, Program Manager	TELEPHONE NUMBER ()
SIGNATURE X	DATE SIGNED 4/12/18

CL 316 (REV. 2/2017) WWW

Figure 59. Sample accident report of AV (3)

Source: DMV

A.5 Appendix E—Shared Mobility Data

A.5.1 Uber Movement FILTERED DATA

Table 16. Origin to All Destination

Origin Movement ID	Origin Display Name	Destination Movement ID	Destination Display Name	Date Range	Mean Travel Time (Seconds)	Range - Lower Bound Travel Time (Seconds)	Range - Upper Bound Travel Time (Seconds)
259	CT 75	1	CT 220.06	1/1/2018 - 1/31/2018, Every day, Daily Average	1570	1309	1882
259	CT 75	6	CT 322.10	1/1/2018 - 1/31/2018, Every day, Daily Average	1614	1357	1919
259	CT 75	9	CT 323.13	1/1/2018 - 1/31/2018, Every day, Daily Average	1339	1106	1619

CT: Census Tract

Source: movement.uber.com

Table 17. Daily time series (evening selected)

Date	Origin Movement ID	Origin Display Name	Destination Movement ID	Destination Display Name	Daily Mean Travel Time (Seconds)	Daily Range - Lower Bound Travel Time (Seconds)	Daily Range - Upper Bound Travel Time (Seconds)	Evening Mean Travel Time (Seconds)	Evening Range - Lower Bound Travel Time (Seconds)	Evening Range - Upper Bound Travel Time (Seconds)
01/01/2018	259	CT 75	259	CT 75	86	56	134	86	56	134
01/03/2018	259	CT 75	259	CT 75	155	69	347	155	69	347
01/04/2018	259	CT 75	259	CT 75	94	51	173	94	51	173

Source: movement.uber.com

Table 18. Chart data (day of week, from 1/1/2018-1/31/2018)

Day of Week	Origin Movement ID	Origin Display Name	Destination Movement ID	Destination Display Name	Date Range	Mean Travel Time (Seconds)	Range - Lower Bound Travel Time (Seconds)	Range - Upper Bound Travel Time (Seconds)
Monday	259	CT 75	259	CT 75	Daily Average	145	72	295
Tuesday	259	CT 75	259	CT 75	Daily Average	134	67	266
Wednesday	259	CT 75	259	CT 75	Daily Average	127	60	265
Thursday	259	CT 75	259	CT 75	Daily Average	155	76	313
Friday	259	CT 75	259	CT 75	Daily Average	178	93	342
Saturday	259	CT 75	259	CT 75	Daily Average	219	105	458
Sunday	259	CT 75	259	CT 75	Daily Average	175	80	383

CT: Census Tract

Source: movement.uber.com

A.5.2 Uber Movement ALL DATA

Table 19. ALL DATA (month aggregate)

sourceid	dstid	month	mean_travel_time	standard_deviation_travel_time	geometric_mean_travel_time	geometric_standard_deviation_travel_time
755	647	3	715.64	383.47	642.25	1.59
755	685	1	1974.41	701.39	1883.32	1.33
746	775	1	1303.87	509.32	1233.46	1.37

Source: movement.uber.com

A.5.3 DiDi Data

Table 20. Raw data of trajectory data from DiDi (city of Xi'an, China, 2016/10/30)

Driver ID	Order ID	Time Stamp	Latitude	Longitude
44a84fc6096312dc5337a91908526384	0498947be30efc2de7af11e6fb01e67a	1477787824	108.91585	34.26921
44a84fc6096312dc5337a91908526384	0498947be30efc2de7af11e6fb01e67a	1477787809	108.91379	34.26921
44a84fc6096312dc5337a91908526384	0498947be30efc2de7af11e6fb01e67a	1477787839	108.91792	34.26921
44a84fc6096312dc5337a91908526384	0498947be30efc2de7af11e6fb01e67a	1477787827	108.91626	34.26921
44a84fc6096312dc5337a91908526384	0498947be30efc2de7af11e6fb01e67a	1477787806	108.91346	34.26922

Source: outreach.didichuxing.com

Table 21. Raw data of order data from DiDi (city of Chengdu, China)

Order ID	Ride Start Time	Ride Stop Time	Pick-up Longitude	Pick-up Latitude	Drop-off Longitude	Drop-off Latitude
mGJrlls.gxjjuafoswAysnom-pwapu8o	1478366285	1478367137	104.07247	30.65341	104.05063	30.69255
nJlhjrf9ttgum5cqowyFjfkf7nmooubq	1478368539	1478369574	104.07502	30.65362	104.0136	30.67191
nJlhjrf9ttgum5cqowyFjfkf7nmooubq	1478368539	1478369574	104.07502	30.65362	104.0136	30.67191
uJzhsm4vBlur4jBpsFHfuiibgmcqnal	1478408628	1478410668	104.112023	30.663959	104.05845	30.64366
qBJrrth1ipmtt7qBtACHfjnq9utekpcm	1478410438	1478412181	104.06401	30.63531	104.04377	30.71717
uBHklop4mykqy_gwouJFtpio6jAluw0d	1478403399	1478405156	104.127065	30.673683	104.07005	30.64377

Source: outreach.didichuxing.com

A.5.4 Kaggle Data

Table 22. Bike sharing demand

datetime	season	holiday	Working day	weather	temp	atemp	humidity	windspeed	casual	registered	count
2011/1/1 0:00	1	0	0	1	9.84	14.395	81	0	3	13	16
2011/1/1 1:00	1	0	0	1	9.02	13.635	80	0	8	32	40
2011/1/1 2:00	1	0	0	1	9.02	13.635	80	0	5	27	32

Source: www.kaggle.com

Appendix References

- Bernardin, V., 2017. The Promise of Big Data—A Case Study. FHWA.
- Hariharan, R., Toyama, K., 2004. Project Lachesis: Parsing and Modeling Location Histories, in: Egenhofer, M.J., Freksa, C., Miller, H.J. (Eds.), *Geographic Information Science, Lecture Notes in Computer Science*. Presented at the International Conference on Geographic Information Science, Springer Berlin Heidelberg, pp. 106–124. https://doi.org/10.1007/978-3-540-30231-5_8
- Jiang, S., 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 2Nd ACM SIGKDD International Workshop on Urban Computing, UrbComp '13*. ACM, New York, NY, USA, pp. 2:1–2:9. <https://doi.org/10.1145/2505821.2505828>
- Longley, P.A., Goodchild, M.F., Maguire, D.J., Rhind, D.W., 2005. *Geographic Information Systems and Science*. John Wiley & Sons.
- Wang, F., Chen, C., 2018. On data processing required to derive mobility patterns from passively-generated mobile phone data. *Transp. Res. Part C Emerg. Technol.* 87, 58–74. <https://doi.org/10.1016/j.trc.2017.12.003>
- Widhalm, P., Yang, Y., Ulm, M., Athavale, S., González, M.C., 2015. Discovering urban activity patterns in cell phone data. *Transportation* 42, 597–623. <https://doi.org/10.1007/s11116-015-9598-x>
- Ye, Y., Zheng, Y., Chen, Y., Feng, J., Xie, X., 2009. Mining Individual Life Pattern Based on Location History. *IEEE*, pp. 1–10. <https://doi.org/10.1109/MDM.2009.11>
- Yin, M., Qiao, S.M., Feygin S, Paiement J.-F, Pozdnoukhov A, 2017. A Generative Model of Urban Activities from Cellular Data, in: *IEEE Transactions in ITS, MobiData '16*. ACM, New York, NY, USA, pp. 25–30. <https://doi.org/10.1145/2935755.2935759>

Americans with Disabilities Act (ADA) Information:

This material can be made available in an alternate format by emailing the Office of Equal Opportunity at wsdotada@wsdot.wa.gov or by calling toll free, 855-362-4ADA(4232). Persons who are deaf or hard of hearing may make a request by calling the Washington State Relay at 711.

Title VI Statement to Public:

It is the Washington State Department of Transportation's (WSDOT) policy to assure that no person shall, on the grounds of race, color, national origin or sex, as provided by Title VI of the Civil Rights Act of 1964, be excluded from participation in, be denied the benefits of, or be otherwise discriminated against under any of its federally funded programs and activities. Any person who believes his/her Title VI protection has been violated, may file a complaint with WSDOT's Office of Equal Opportunity (OEO). For additional information regarding Title VI complaint procedures and/or information regarding our non-discrimination obligations, please contact OEO's Title VI Coordinator at (360) 705-7082.
