ABSTRACT

Title of Thesis:	MULTIMODAL TRAVEL MODE IMPUTATION BASED ON PASSIVELY COLLECTED MOBILE DEVICE LOCATION DATA
	Mofeng Yang, Master of Science, 2020
Thesis Directed By:	Professor, Lei Zhang, Department of Civil and Environmental Engineering

Passively collected mobile device location (PCMDL) data contains abundant travel behavior information to support travel demand analysis. Compared to traditional travel surveys, PCMDL data have larger spatial, temporal and population coverage while lack of ground truth information. This study proposes a framework to identify trip ends and impute travel modes from the PCMDL data. The proposed framework firstly identify trip ends using the Spatiotemporal Density-based Spatial Clustering of Applications with Noise (ST-DBSCAN) algorithm. Then three types of features are extracted for each trip to impute travel modes using machine learning methods. A PCMDL dataset with ground truth information is used to calibrate and validate the proposed framework, resulting in 95% accuracy in identifying trip ends and 93% accuracy in imputing five travel modes using the Random Forest (RF) classifier. The proposed framework is then applied to two large-scale PCMDL datasets, covering Maryland and the entire U.S. The mode share results are compared against travel surveys at different geographic levels.

MULTIMODAL TRAVEL MODE IMPUTATION BASED ON PASSIVELY COLLECTED MOBILE DEVICE LOCATION DATA

by

Mofeng Yang

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Master of Science
2020

Advisory Committee:
Professor Lei Zhang, Chair
Professor Paul Schonfeld
Professor Taylor Oshan

Dedication

To my beloved parents Kun Yang and Peifan Li, and my girlfriend Zhiyue Xia.

Acknowledgements

This research was partially funded by Federal Highway Administration (FHWA). Opinions herein do not necessarily represent the views of the research sponsors. The author is responsible for the statements in the thesis.

First, I would like to express my sincere gratitude to my advisor, Dr. Lei Zhang for his continuous support and guidance during the past two years. Dr. Zhang offered me the opportunity to study and work at the University of Maryland with financial support through a research assistantship, which enabled me to reach where I am now. He has always been my role model that I am willing to be in my future career.

I would also like to thank Dr. Paul Schonfeld and Dr. Taylor Oshan for serving on my master thesis committee and offering me their valuable comments. I am extremely grateful for Dr.Schonfeld for his guidance among the courses I took in my first year. I would also like to express my special gratitude to Dr. Oshan for the perfect course I took with him at the Department of Geographical Science.

I also want to thank my colleagues at Maryland Transportation Institute that helped me on this thesis: Sepehr Ghader, Aref Darzi, Yixuan Pan, Jun Zhao, Weiyi Zhou and Minha Lee. Thank you all for the help and advises.

Last, I feel also thankful for my girlfriend, Zhiyue Xia, who always supports me whenever I met obstacles. And I would like to thank my parents, Kun Yang and Peifan Li, for always respecting my opinions and decisions.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AADT	Annual Average Daily Traffic
ANN	Artificial Neural Networks
AWS	Amazon Web Service
BMC	Baltimore Metropolitan Council
BTS	Bureau of Transportation Statistics
CASI	Computer-Assisted Self-Interview
CATI	Computer-Assisted Telephone Interview
CBSA	Core-based Statistical Area
CDR	Call Detail Record
DBSCAN	Density-based Spatial Clustering of Applications with Noise
DNN	Deep Neural Networks
DMV	Washington Metropolitan Area
FHWA	Federal Highway Administration
GPS	Global Positioning System
HPMS	Highway Performance Monitoring System
KNN	K-Nearest Neighbors
LBS	Location-based Service
MWCOG	Metropolitan Washington Council of Government
NHTS	National Household Travel Survey
NTM	National Transit Map
OD	Origin and Destination
PAPI	Paper-And-Pencil Interview
PCMDL	Passively Collected Mobile Device Location
SMOTE	Synthetic Minority Over-sampling Technique
SVC	Support Vector Classifier
SVM	Support Vector Machine
TAZ	Traffic Analysis Zone
TPB	Transportation Planning Board
UMD	University of Maryland
U.S.	the United States
USDOE	the United States Department of Energy
USDOT	the United States Department of Transportation
XGB	eXtreme Gradient Boosting

Chapter 1: Introduction

1.1 Background

Understanding travel behavior has always been one of the most important tasks in the realm of transportation planning. An accurate measurement of travel behavior can help governments and agencies understand how it evolves and better allocate resources in support of different transportation planning applications.

Traditionally, researchers and practitioners design and conduct travel surveys to obtain household and individual travel behavior data, including trip origins and destinations, trip distance, trip time, trip purposes, travel modes, etc. Some of the most famous travel surveys conducted in the United States (U.S.) include the National Household Travel Survey (NHTS) [1], American Travel Survey [2], etc. However, traditional travel surveys require complex planning, design, large human labor and costs to obtain reasonable estimates from samples to the population level. For instance, the average cost of a travel survey is estimated at $487,000 implying about $9.7 million annually in the U.S. [3].

In the past two decades, along with the technology advancement in mobile sensors and mobile networks, passively collected mobile device location data, PCMDL data in abbreviate, has been growing drastically in terms of data coverage and data size. In the realm of transportation, the abundant personal movement information in the PCMDL data has great potentials to help researchers and practitioners understand the whole picture of human travel. Compared to traditional travel surveys, it has larger spatial,

temporal and population coverage while lack of ground truth, such as trip origins and destinations, trip purpose and travel modes. The missing information should be imputed using appropriate methods with additional data inputs to extract useful travel behavior data. In addition, though promising, the sources of PCMDL data can be various, including Global Positioning Service (GPS) devices, cellular network, Bluetooth, Wi-Fi, etc. The similarity and the difference between different PCMDL data sources are also one important thing that should be taken care of.

1.2 Research Objective

The objective of this study is to develop a methodological framework to obtain travel behavior information from the PCMDL data by identifying trips and imputing travel mode. The identified trips should include accurate trip origins and destinations, trip start time, and trip time information. The imputed travel modes should include multimodal travels, including drive, bus, rail and non-motorize travel modes. In order to fulfill the research objective, four tasks are identified as shown below: (1) evaluating the state-of-the-practice applications and the state-of-the-art methods based on PCMDL data and identifying the key research gap; (2) developing a suitable algorithm to extract accurate trip ends from the PCMDL data; (3) exploring what are the important features that can be used to impute travel modes based on PCMDL data and other publicly available information; (4) validating the travel mode imputation results using the traditional travel survey data and other publicly available data sources.

1.3 Research Contribution

The main contribution of this study can be classified into three folds: (1) Uniqueness of the Data; (2) Methodology; (3) Comprehensive comparison process.

1.3.1 Uniqueness of the Data

This is the first study that utilizes three PCMDL datasets from different sources with various spatial and temporal coverages. The first PCMDL dataset is collected from one of the most advanced Mobility-as-a-Service (MaaS) mobile application, incenTrip. This dataset has travel behavior data with ground truth information including trip origin and destination, travel modes with user recall. The intermediate locations of the trip are also recorded. The second and the third PCMDL dataset is obtained from one of the leading PCMDL data vendors, covering Maryland and its peripheral with a temporal coverage of one day and the entire U.S. with a temporal coverage of seven days respectively.

Apart from the PCMDL data, this study also utilized the public available multimodal transportation networks and stations information collected from the United States Department of Transportation (USDOT) Bureau of Transportation Statistics (BTS) National Transit Map (NTM) [4], which contains bus and rail (metro included) networks and stations information.

1.3.2 Methodology and Comparison

This study examines the state-of-the-practice applications and state-of-art-methods on processing the PCMDL data. Based on the literature review result, a new framework is

proposed to process the PCMDL data from raw location points into trips with imputed travel modes. The proposed framework has two parts: the first part utilizes a Spatiotemporal Density-based Spatial Clustering of Applications with Noise (ST-DBSCAN) algorithm to identify the activity locations with only PCMDL data information; the second part first construct features from the identified trips and examines the performance of various machine learning methods. The Random Forest (RF) algorithm is identified as the best method for travel mode imputation in terms of prediction accuracy, which has the capability to produce the best accuracy and prevents the overfitting problem.

1.3.1 Application Potential

The methodological framework proposed in this study is developed using a real-world dataset and applied on two other PCMDL datasets for case study purposes. The proposed framework is compared to be general at different geographies. The additional data used to support the proposed framework is publicly available and can be generalized to an even larger population.

1.4 Research Approach and Outline

The research approach of this study starts with a comprehensive literature review of the state-of-the-practice applications and the state-of-the-art methods about the PCMDL data. The key research gap is identified from the literature review. Then, three PCMDL datasets used in this study are introduced. With the datasets introduced, a methodological framework is proposed and applied to the datasets. The results are further compared with the travel surveys and other publicly available information.

The outline of this thesis is organized as shown in Figure 1-1. Chapter 2 first categorizes the PCMDL data into four categories by providing a comprehensive literature review about the state-of-the-practice applications using PCMDL data, and the state-of-the-art methodologies applied to the PCMDL data. Chapter 3 introduces the three PCMDL datasets used in this study, which includes one dataset from an active mobile application and two other datasets from one of the leading PCMDL data vendors in the world. The multimodal transportation network data and the travel survey data are also introduced. Chapter 4 demonstrates the proposed methodological framework to identify trip ends and impute travel modes from the PCMDL data. Several machine learning methods are briefly introduced. The feature construction process, model training, model selection and model applications are also included in this section. Chapter 5 first shows the calibration and comparison of the proposed methodological framework using the first PCMDL datasets with ground truth information. Then, two case studies further validate the framework using two PCMDL datasets. The detailed comparison process is also incorporated in this chapter. Finally, Chapter 6 summarizes the conclusion and suggests future research directions.

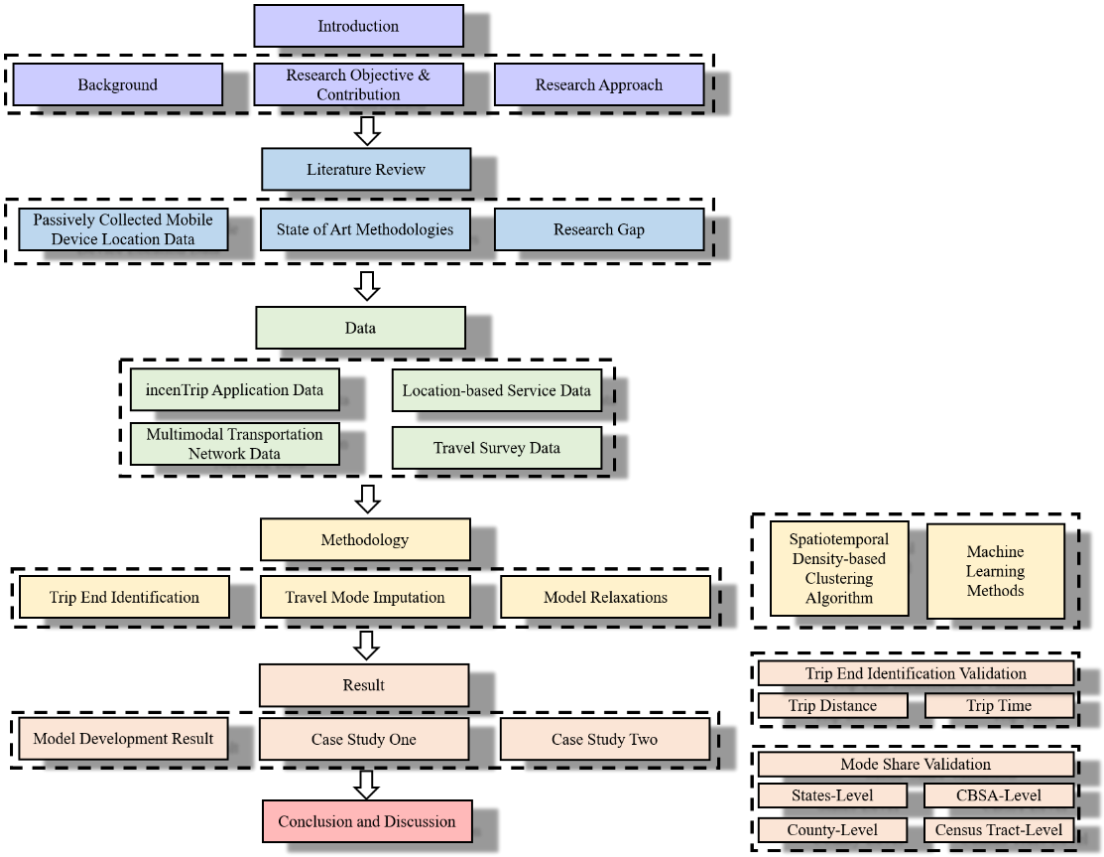


Figure 1-1. Thesis Outline.

Chapter 2: Literature Review

2.1 Passively Collected Mobile Device Location Data

2.1.1 GPS-Enhanced Travel Survey Data with User Recall

Travel survey serves as an important tool to obtain person-level or household-level travel behavior pattern, supporting both traditional four-step and activity-based travel demand model in the regional transportation planning process [5]. Traditional methods to conduct travel surveys usually require respondents to record their daily trips with original paper-and-pencil interview (PAPI), computer-assisted telephone interview (CATI), and computer-assisted-self-interview (CASI) [6,7]. However, these methods are prone to several well-known biases, such as under-reported trips, inaccurate travel times, and travel distances [8,9].

Since the late 1990s, with the commercialization of the Global Positional System (GPS), GPS data logger was also introduced to enhance the quality of travel surveys with personal longitudinal location data. Table 1 summarized the location data-enhanced travel surveys conducted in the U.S. At initial stages, the GPS data logger was installed in the vehicle and charged by the vehicle battery [10–18]. It records location data seconds by seconds when the vehicle is moving and would stop recording data when the vehicle is not moving, for example, if the vehicle speed falls below five miles per hour for a continuous period of thirty minutes or more [13]. This approach was proved to be effective, but it only captured respondents' vehicle trips. Later on, the wearable GPS [19–22] further allowed respondents to carry within the bag such that

trips traveled by other non-vehicle travel modes could also be obtained. The drawback of the wearable GPS data logger was that it needed to be changed frequently. Some surveys utilized both in-vehicle and wearable GPS data loggers to take advantage of both devices [23–25].

Table 2-1. Literature Review of GPS-Enhanced Travel Survey in the U.S.

Data Collection Methods	Region	Year
In-Vehicle GPS Logger	Lexington [10]	1997
	California [11]	2001
	Kansas City [12]	2004
	Austin/San Antonio [13]	2006
	Metropolitan Washington [14]	2007-08
	Metropolitan Baltimore [14]	2007-08
	Houston-Galveston [15]	2008-09
	El Paso [16]	2010-11
	Wichita Falls [17]	2010-11
	Abilene [18]	2010-11
Wearable GPS Logger	Minneapolis – St. Paul [19]	2010
	Delaware Valley [20]	2012-13
	New Mexico [21]	2013
	Nevada [22]	2014
In-Vehicle and Wearable Combined	Chicago [23]	2007
	Atlanta [24]	2011
	California [25]	2010-12
Smartphone Applications	Puget Sound [26–28]	2014, 2015, 2017
	Madison County [29]	2015

Nowadays, the advances of mobile device location data provide an alternative way to conduct travel surveys by using smartphones to passively collect respondent's location data. For instance, the most recent smartphone-enhanced travel surveys conducted in the United States enable the smartphone App running in the background to passively collect location data continuously with a fixed interval (usually 1 second) [26–29]. Also, in New Zealand, a smartphone-based system for personal travel survey was proposed and tested in real-world implementation [30]. This type of smartphone location data is usually generated by the location service providers not only using GPS,

but also Bluetooth (where those are available) along with crowd-sourced Wi-Fi hotspot and cell tower locations to determine the device's approximate location.

2.1.2 GPS Data with No User Recall

Another type of GPS data is also passively collected but without any user recall. This type of data is widely collected with the in-vehicle GPS device for both passenger vehicles and trucks. For instance, INRIX Traffic [31] as a data provider collects GPS probe data from commercial vehicle fleets, connected vehicles and mobile apps. The data is further aggregated into link or corridor level to provide a real-time estimation of traffic speed and travel time [32-34].

2.1.3 Cellular Data

Two types of data are included in the cellular data: Call Detail Record (CDR) and sightings. Call Detail Record (CDR) data is generated when a phone communicates with the cell tower in the cellular network, for instance when a phone call or text message is made by the phone. The location information of CDR data is the cell tower locations thus it fully depends on the density of the cellular network and does not reflect the actual location of the device [35]. Another type of cellular data is called sightings, where the location information is calculated via triangular calculation with several cell towers [35]. Both types of cellular data has been widely used in studying human mobility patterns in the past two decades [36-38].

2.1.3 Location-based Service Data

Similar to cellular data, the Location-based Service (LBS) data is generated when a smartphone updates the App periodically with the best location accuracy, based on the currently-available location providers such as Wi-Fi, Bluetooth, cellular tower and GPS [35,39]. The LBS data can reflect the exact location of the device and thus providing invaluable location information that describes depict person-level mobility pattern. Also, in most cases, the LBS data has a higher spatial precision and sample rate than the CDR data [35-40].

The most recent research proposed a "Divide, Conquer and Integrate" (DCI) framework to process the LBS data to extract mobility patterns in the Puget Sound region, and the result was aggregated at census tract-level and compared with household travel survey [40]. The DCI framework was also applied to another LBS dataset in Texas to analyze the impact of hurricane Harvey on travel patterns [41].

In the industry, many location-intelligence companies have started to deliver products using the LBS data. For instance, StreetLight Data Inc. produced the Annual Average Daily Traffic (AADT) estimates, Bike and pedestrian analysis, etc. with the large-scale LBS data for the entire United States purchased from the data vendors [42]; AirSage leveraged LBS data to develop a traffic platform that can estimate traffic flow, speed, congestion and road user sociodemographic for every road and time of day [43].

2.1.4 Summary of PCMDL Data

In summary, these four types of PCMDL data are different in terms of spatial coverage, temporal coverage, population coverage, sample rate. Table 2-2 summarize the overall comparisons between these four types of PCMDL data.

Table 2-2. Comparisons among PCMDL data.

Data	Spatial Coverage	Temporal Coverage	Population Coverage	Sample Rate
GPS Data with User Recall	Low	Low	Low	High
GPS Data with No User Recall	Medium	High	Low	High
Cellular Data	High	High	High	Low
LBS Data	High	High	High	Low

The GPS data with user recall, which is usually collected for travel survey purposes, has the highest sample rate (usually 1 second) that provides second-to-second trajectories with respondents' confirmed ground truth information. However, the limitation is that the travel surveys usually sample a small percentage of respondents in small regions with a short survey period, resulting in low spatial, temporal and population coverage. Thus, this type of data cannot reflect population-level travel behavior. Additional weighting processes need to be combined to provide a statistically reasonable result.

The GPS data with no user recall usually has the same level of sample rate as the GPS data with user recall. Though with low population coverage, the spatial coverage and temporal coverage is improved.

The sample rate of cellular data and LBS data are solely based on mobile device users' frequency on using either telecommunication or location-based services. However, since a large proportion of the population owns a mobile device, cellular data and LBS data have significantly higher spatial, temporal and population coverage over the other types of data, while the ground truth information is missing.

2.2 Extracting Trips from Passively Collected Mobile Device Location Data: State of Art Methodologies

2.2.1 Trip End Identification

The trip end identification algorithm for GPS data with high frequency has been well-developed and used in practical application. To obtain accurate trip ends, the traditional way is the rule-based trip end identification methods. This type of method designs rules and parameters. The trip ends are obtained by applying the rules to location data point by point and at the same time examining the intra-relationship between several consecutive location points. Though proven to be effective, the rule-based methods highly rely on the design of the rule and the corresponding rule-based parameters. Most rules are complexly designed and the physical meanings behind is hard to interpret. Also, the parameters used in these rules are mostly defined by domain knowledge, such as dwell time, speed, etc. [6, 44-54]. In recent years, some researchers also leveraged the supervised machine learning method as a supplement to the rule-based methods, which classified each location point as static or moving [55-57]. However, the complexity of designing the rule and the un-interpretable result problems still exist.

Different clustering methods were also applied to obtain trip ends by identifying people's potential activity places from the location data [58-61]. The most recent one in the literature utilized the Spatio-temporal clustering method with three combined optimization models to detect trip ends [61]. In their study, the respondents were still asked to record the trip starts and trip ends in the smartphone application. Moreover, the proposed trip end identification method still largely relied on the speed attribute, which could not always be available or observed accurately due to the heterogeneity of different types of smartphones.

In recent years, there's also a special focus on deriving the trip ends from LBS data. In [40], a Divide, Conquer and Integrate (DCI) framework was proposed to process the LBS data to extract mobility patterns in the Puget Sound region. The proposed framework combined a rule-based method and incremental clustering method to handle the bi-modal distributed LBS data.

2.2.2 Travel Mode Imputation

Travel mode imputation can be categorized into mainly two approaches: (1) trip-based approach; and (2) segment-based approach. The trip-based approach is based on the already identified trip ends, where each trip has only one travel mode to be imputed. The segment-based approach separates the trip into a fixed-length segment (time or distance) and then impute the travel mode for each segment. Then the segment with the same travel mode will be further merged to form a single-mode trip. This study considers the trip-based approach mainly. Table 2-3 summarizes some typical methods for travel mode imputation using the trip-based approach.

Table 2-3. Literature Review on Travel Mode Imputation.

Authors	Sample Rate	Model	Features	Modes	Accuracy
Gong et al. [52]	/	Rules	Speed, Acceleration, Transit Stations, Transit Network	D,T,Bu,W,S,	82.6%
Stenneth et al. [62]	30 s	RF	Speed, Acceleration, Heading change, Bus location, Transit Network	D,Bu,Tr,Bi,W,St	93.7%
Bruunauer et al. [63]	1-10 s	MLP	Speed, Acceleration, Bendiness	D,Bu,Tr,Bi,W	92%
Xiao et al [64]	1s	BN	Speed, Acceleration, Trip Distance	D,Bu,W,Bi,Eb	92%
Broach et al [65]	4s	MNL	Speed, Acceleration	D,T,W,Bi	96%
Shafique and Hato. [66]	0.1s	RF	Resultant acceleration	D,Bu,W,Bi,Tr,S	99.8%
Wang et al. [67]	1s	RF	Speed, Acceleration, Orientation, Distance/Duration, Sociodemographic	D,Bu,W,Bi,Eb	93.1%

** D: Drive; Bu: Bus; Tr: Train; W: Walk; Bi: Bike; S: Subway; Eb: Electric-Bike; St: Stationary; RF: Random forest; MLP: Multilayer perceptron; BN: Bayesian network; MNL: Multinomial logit model.*

It can be observed that some typical features used are speed (average, minimum, maximum, and quantiles), acceleration (average, minimum, maximum, and quantiles) [52,62-67]. Specifically, when the sample rate is higher than 10 seconds, the speed and acceleration features are more important to separate between different travel modes, which can be imputed solely by the data itself. However, as mentioned in [62], the higher the sample rate, the more battery the mobile device will need to consume. To maintain the same level of imputation accuracy and at the same time reduce battery consumption, additional features are added such as real-time transit information [62], multimodal transportation network [52,62], sociodemographic information [67] etc.

2.3 Research Gap

Both the state-of-the-practice applications and the state-of-the-art methodologies can accurately identify trip ends and impute travel modes based on high sample rate GPS data with ground truth information. However, these methods have neither been applied nor compared on the emerging PCMDL datasets, LBS data in particular.

Also from the state-of-the-practice application side, though location-intelligence companies like StreetLight Data [42] and AirSage [43] have developed multimodal transportation analysis products based on LBS data, their trip ends and travel mode imputation results have never been compared yet.

The key research gap identified from the literature review indicates that few studies have focused on developing methods and comparison processes for extracting travel behavior data from LBS data. This study aims to fill this gap by proposing a methodological framework that developed based on GPS data and then applying it on two large-scale LBS datasets. A comparison process is also included in the framework to comprehensively validate the result against the travel surveys.

Chapter 3: Data

3.1 GPS Data with User Recall: incenTrip Mobile Application

3.1.1 incenTrip Introduction

incenTrip (*incentrip.org*), was developed by National Transportation Center (NTC) at the University of Maryland (UMD) for the "Integrated, Personalized, Real-time Traveler Information and Incentive" (iPretii) project, funded by U.S. Department of Energy's (DOE) Advanced Research Projects Agency-Energy (ARPA-E). The incenTrip application was officially launched on Aug 28th, 2019 with the initial support of the Metropolitan Washington Council of Government (MWCOG). Since then, incenTrip has been incentivizing users for taking transit, multimodal or non-motorized travel modes. The corresponding trip data with travel mode imputed and confirmed by users is collected in order to provide corresponding incentives to nudge behavior changes.

The incenTrip application would obtain location data from Google Maps API with a pre-defined sample rate and store the data in the Amazon Web Services (AWS) with privacy protection. The proposed framework in this study would then be applied to identify each user's trips from the raw location data and then update the identified trips back into the database. Last, the application would show the trips for each user on the mobile application page and let users recall and confirm the trips and travel modes they made before.

The service area of incenTrip covers the entire Washington Metropolitan Area (DMV) and the Baltimore Metropolitan Council Area (BMC), as shown in Figure 3-1. This service area covers all kinds of daily travel modes, including metro lines, light rail lines, commuter rail lines (MARC), numerous bus lines and capital bike share stations.

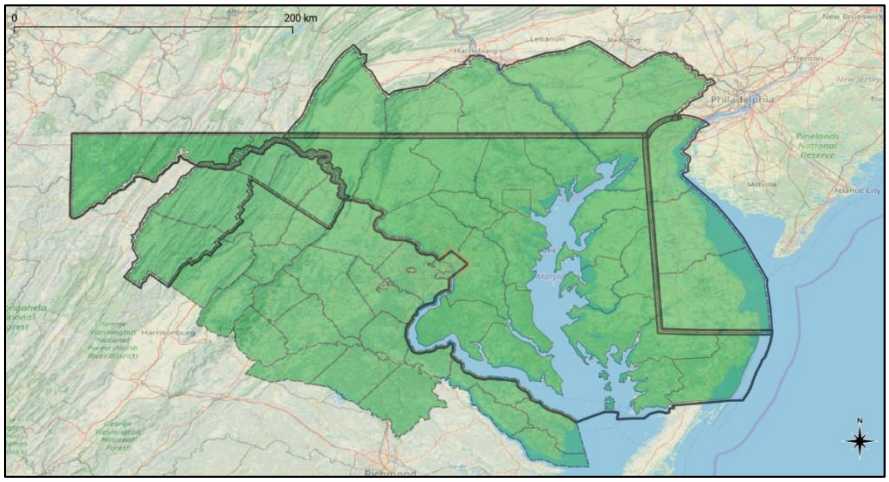


Figure 3-1. incenTrip Service Area.

3.1.2 Data Description

Two datasets from incenTrip are collected to calibrate and validate the proposed trip end identification algorithm and travel mode imputation model respectively. Table 3-1 summarizes the two datasets.

Table 3-1. Summary of the Two Datasets.

Sample Rate (moving / not moving)	Duration	Number of Reported Trips by Travel Mode				
		Drive	Bus	Rail	Bike	Walk
Dataset One used for Trip End Identification						
1s / 1s	03/13 - 03/15	12	3	6	0	2
2s / 5s	02/21 - 04/29	427	123	84	9	72
5s / 10s	04/13 - 05/02	116	192	76	20	37
5s / 30s	05/03 - 05/16	55	101	64	40	21
15s / 30s	05/21 - 06/12	95	130	85	16	2
Total		705	550	315	85	134
Dataset Two used for Travel Mode Imputation						
/		6,064	1,403	1,824	1,496	1,901

The first dataset was collected from March to June 2019, when fifteen testers were hired to install the test version of incenTrip application, most of them being graduate students and faculties from the University of Maryland. For each week, each tester was assigned to an area to travel with different travel modes. At the same time, all of them were required to record detailed information for each trip, including the start date, start time, end date, end time, origin street address, destination street address, travel time and travel mode. Five phases of testing were conducted with different sample rates. At the beginning of each phase, testers were asked to install a new version of the incenTrip application with the new sample rate. The major consideration of setting different sample rates was to reduce the impact of the smartphone application on smartphone battery draining speed and at the same time ensure all the travel information was collected. We also tested sample rate including both lower and higher than five seconds. It should be noted that for the 1s/1s sample rate, only two testers were involved for two days since the battery consumption was too large.

The second dataset was collected from March 2019 to January 2020, including the first dataset. The trips with the travel modes confirmed by users are extracted from the AWS database. For trips with correctly imputed travel mode, the corresponding data are directly extracted. For trips with wrongly imputed travel modes, the corresponding data are extracted and labeled with the users' corrected travel mode label. This dataset does not divide trips into different sample rates (most of the trips have a sample rate of 15s/30s) since the sample rate is an important feature that is stochastic in the LBS data.

3.2 Location-based Service Data

The LBS data is obtained from one of the leading data vendors in the U.S, including the whole year of 2017, 2018. The real-time data is also available upon requests.

In this study, two small LBS datasets are sampled with different spatial, temporal and population coverages. Table 3-2 summarizes the descriptive statistics of these two datasets. Dataset A covers the entire state of Maryland, Washington D.C. and part of northern Virginia. It has a temporal coverage of one day on September 12nd in 2017, including 474,634 devices. Dataset B covers the entire U.S. It has a temporal coverage of seven days from August 1st to August 7th in 2017. 1% of the total number of devices observed is sampled via random draw and used for this study, including 266,149 devices.

Table 3-2. LBS Data Descriptive Statistics.

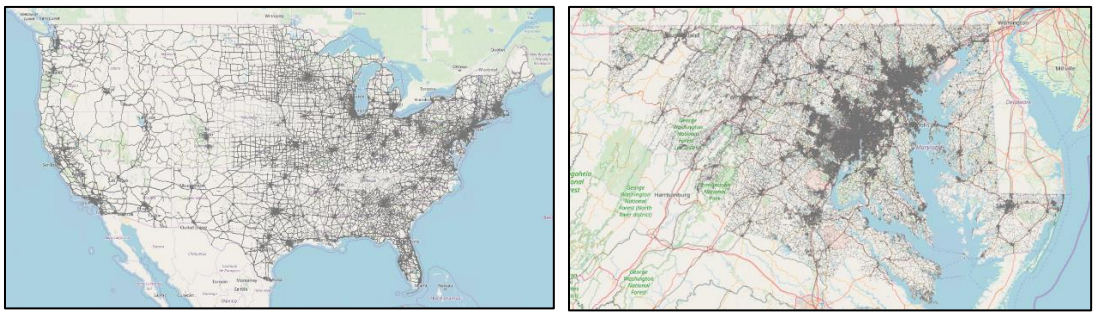
Dataset	Spatial Coverage	Temporal Coverage	Sample Rate	Number of Device
A	Maryland and it's peripheral	1 day	~100%	474,634
B	the United States	7 days	~3%	266,149

The major consideration of selection of these two datasets is the computing time. For dataset A, it has small temporal-spatial and temporal coverage while capturing 100% devices that observed in the region. For dataset B, since the spatial coverage is expanded to the entire U.S. and the temporal coverage is expanded to 7 days, to reduce the total computing time, only 3% of the device observed in the U.S. over 7 days are captured to have a comparable data size with dataset A.

3.3 Multimodal Transportation Network Data

This study also collects the multimodal transportation network data including drive, bus, rail networks and bus stop locations in order to construct features that will be fed into the travel mode imputation models.

Two different drive networks are collected. The first network is collected from Highway Performance Monitoring System (HPMS) [68] that covers the entire U.S. including national freeway and arterial road networks. The second network is collected from HERE [69], a mapping and location data and related service provider, including all types of roads in the state of Maryland and its peripherals. Figure 3-2 illustrates the road networks.



(a) (b)

Figure 3-2. Road Network: (a) national drive network; (b) Maryland drive network.

The national bus and rail network, and the bus stops data are collected from the United States Department of Transportation (USDOT) Bureau of Transportation Statistics (BTS) National Transit Map (NTM) [4]. Figure 3-3 shows the multimodal networks.

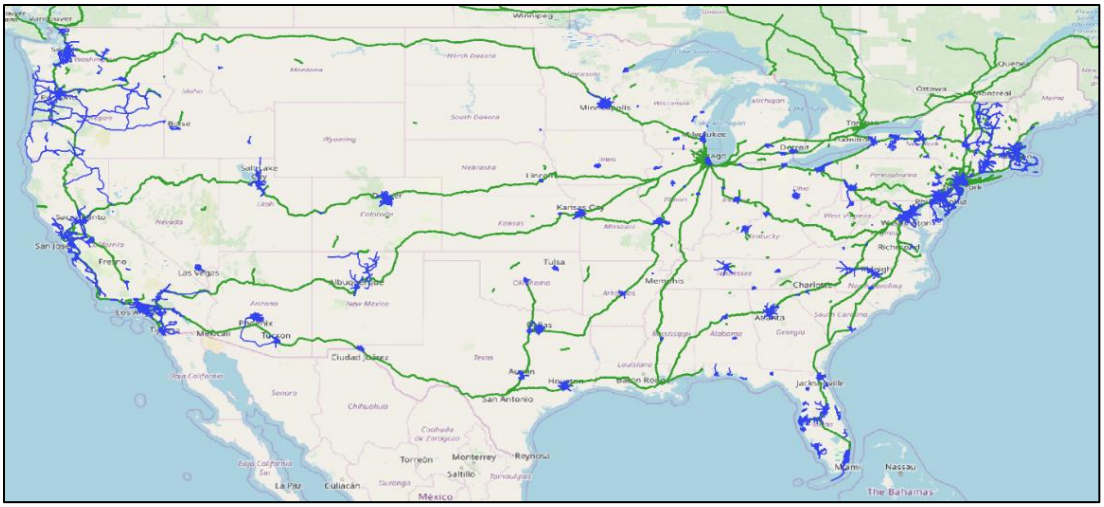


Figure 3-3. Multimodal Transportation Network

**Green represents rail network, blue represents bus network*

3.4 2017 National Household Travel Survey

The National Household Travel Survey (NHTS) [1] is a national-level travel survey conducted by USDOT Federal Highway Administration (FHWA), collecting travel behavior data by U.S. residents in all 50 states and District of Columbia, including trip origin and destinations, trip time, trip purposes (work, school, other) and travel modes (private vehicle, public transportation, pedestrian and cycling).

The 2017 NHTS required that every household member age 16 and older complete a retrieval interview for the household to be considered complete, finally with a total number of 129,696 household data collected. The survey sample data is used to develop household, person, trip and vehicle weights separately in order to produce the population-level travel statistics [70].

In this study, the trip distance and trip time distribution from the 2017 NHTS are used to validate the trip end identification results from the two LBS dataset. The imputed

mode shares at different geographic levels are also compared to the mode share from 2017 NHTS.

3.5 2007/2008 TPB-BMC Household Travel Survey

The 2007/2008 TPB-BMC Household Travel Survey (HHTS) is conducted by the Transportation Planning Board (TPB) in Baltimore and Washington regions from February 2007 to March 2008 using the same survey designs [71,72]. This survey covered nearly 14,000 households and can provide mode share information at Traffic Analysis Zone (TAZ) level. In this study, the mode shares for nine counties aggregated from TPB-BMC HHTS are used to validate the imputed mode shares from the second LBS dataset.

Chapter 4: Methodology

4.1 Methodological Framework

The proposed methodological framework of this study is shown in Figure 4-1, including two major parts: model development and model application. In the model development part, firstly, the first dataset collected from the incenTrip application is used to calibrate and validate the trip end identification algorithm. Then the second dataset collected from the incenTrip application is used to train and validate the travel mode imputation model.

Then, before the model applications, the similarity and difference between the incenTrip data and the LBS data is discussed in order to relax the constraints of the developed models. After that in the model application part, the relaxed models are directly applied to the two LBS datasets as mentioned in previous sections and the results are compared against the travel surveys

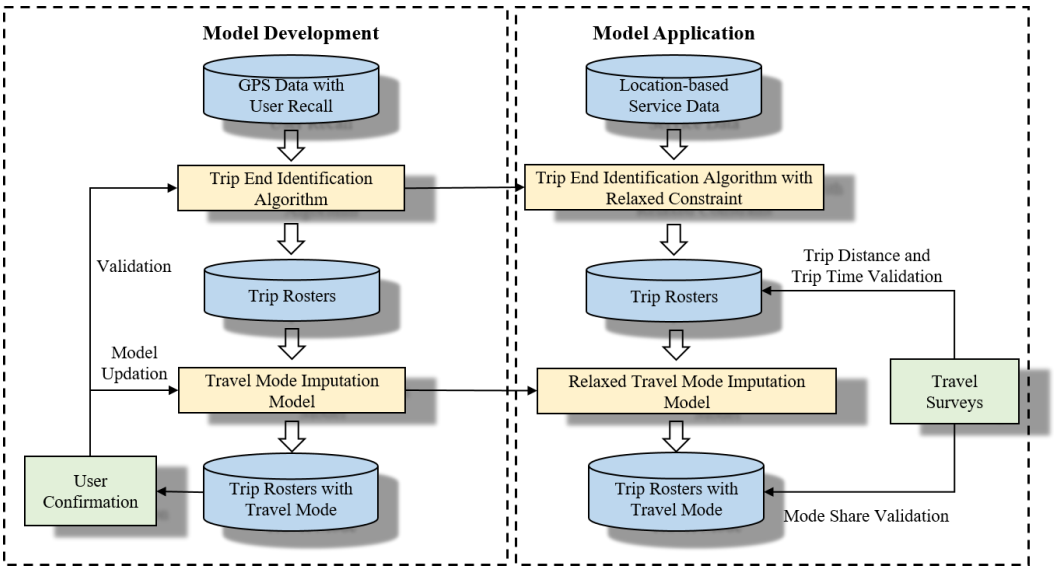


Figure 4-1. Methodological Framework

4.2 Trip End Identification

4.2.1 Potential Activity Location Identification

Considering a person's daily trajectory, it's very common that he or she makes multiple stops at different places each day. In this study, as illustrated in Figure 4-2, the stops were categorized into two categories, namely Activity Stop (AS) and Non-Activity Stop (NAS). AS represents a stop where an actual activity takes place, such as home, workplace, restaurant, shopping mall, etc. NAS represents a stop where no activity takes place or the activity takes a very short amount of time, usually including stopping at a traffic light, picking up people within a short range of time, etc. In this study, only ASs were considered as actual trip ends and the trajectory between two consecutive ASs were considered as an actual trip.

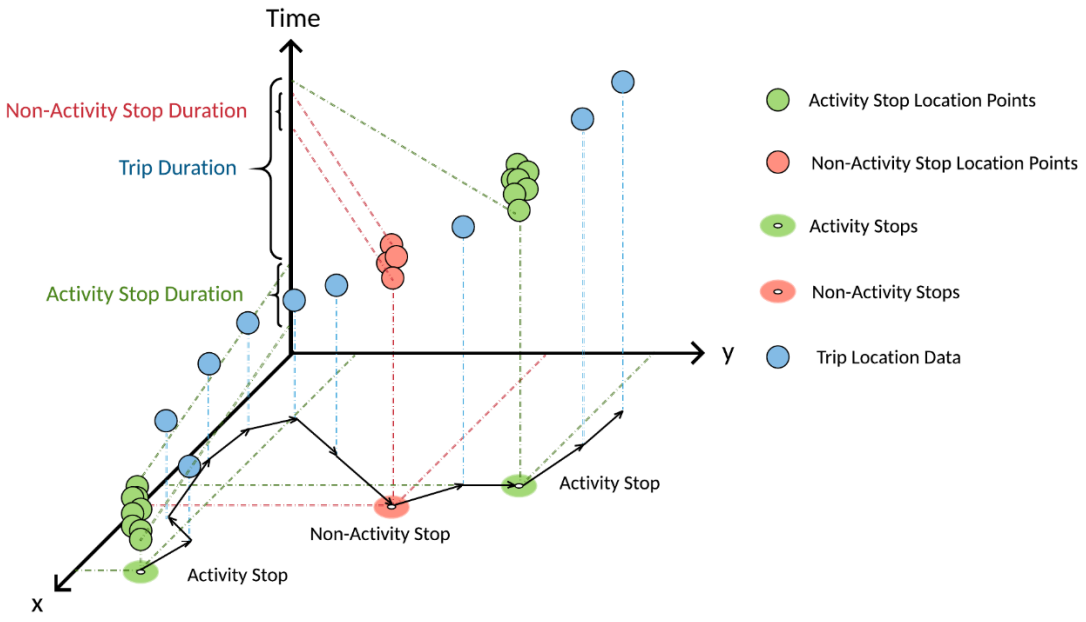


Figure 4-2. Illustration of a Person's Trajectory.

To identify all potential ASs, a Spatiotemporal Density-based Spatial Clustering of Applications with Noise (ST-DBSCAN) [73] was applied to fit the data. The original ST-DBSCAN is an extended version of the traditional DBSCAN algorithm [74] with consideration of both spatial and temporal constraints, as illustrated in Figure 4-3. The temporal constraint was able to handle the scenario when a person stays at the same place multiple times per day, such as going out for lunch and return to the office, going back home, etc. Below shows a short description of the ST-DBSCAN algorithm, detailed information can be found in [73,74]

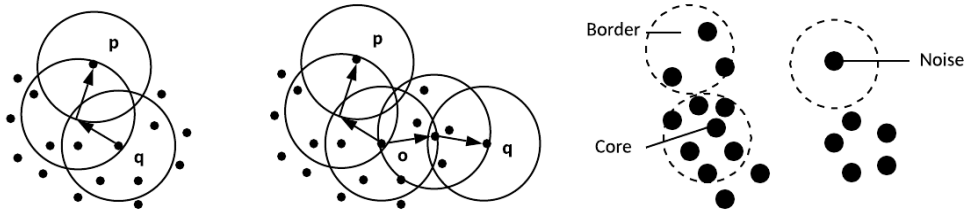


Figure 4-3. Illustration of DBSCAN Algorithm.

Definition 1 (*Clustering*). Given a database of *n* data objects $D = \{o_1, o_2, ..., o_n\}$. The process of partitioning D into $C = \{C_1, C_2, ..., C_k\}$ the base on a certain similarity measure is called clustering, C_i's are called clusters, where $C_i \subseteq D, (i = 1, 2, ..., k), \bigcap_{i=1}^{k} C_i = \varnothing$ and $\bigcup_{i=1}^{k} C_i = D$.

Definition 2 (*Neighborhood*). It is determined by a distance function (e.g., Manhattan Distance, Euclidean Distance) for two points p and q, denoted by $\text{dist}(p, q)$.

Definition 3 (*Eps-neighborhood*). The Eps-neighborhood of a point p is defined by $\{q \in D \mid \text{dist}(p, q) \leq \text{Eps}\}$.

Definition 4 (*Core object*). A core object refers to such a point that its neighborhood of a given radius (Eps) has to contain at least a minimum number (MinPts) of other points.

Definition 5 (*Directly density-reachable*). An object *p* is directly density-reachable from the object *q* if *p* is within Eps-neighborhood of *q*, and *q* is a core object.

Definition 6 (*Density-reachable*). An object p is density-reachable from the object q with respect to Eps and MinPts if a chain of object $p_1,...,p_n$, $p_1 = q$ and $p_n = q$ such that p_{i+1} is directly density-reachable from pi with respect to Eps and MinPts, for $1 \le i \le n$, $p_i \in D$.

Definition 7 (*Density-connected*). An object *p* is density-connected from the object *q* with respect to Eps and MinPts if an object $o \in D$ such that both p and q are density-reachable from *o* with respect to Eps and MinPts.

Definition 8 (*Density-based cluster*). A cluster C is a non-empty subset of D satisfying the following "maximality" and "connectivity" requirements:

$\forall p,q$: if $q \in C$ and *p* is density-reachable from *q* with respect to *Eps* and *MinPts*,

then $p \in C$

$\forall p,q \in C$: *p* is density-connected to *q* with respect to *Eps* and *MinPts.*

Definition 9 (*Border object*). An object p is a border project if it is not a core object but density reachable from another core object.

Definition 10 (Noise). An object p is a border project if it is not a core object but density reachable from another core object.

Three thresholds are demonstrated for the ST-DBSCAN used in this study: (1) the spatial threshold *s* represents the distance falling within the activity distance range, calculated from a geographic distance; (2) the temporal threshold *t* represents the minimum duration of an activity, defined by a given value. (3) the minimum neighbor's *m* threshold represents the density of the location points. After all the potential ASs were identified, we considered that a trip end is the first stopped point of a cluster and a trip start is the point immediately following the last stopped point of a cluster [52]. Thus, all the location points between a trip start and trip end can be labeled as waypoints of a trip.

4.2.2 Activity Locations Extraction and Non-Activity Locations Elimination

With all potential trips identified, the second step was to distinguish between ASs and NASs. Two parameters were proposed, *s_act*: maximum activity distance range and *t_act*: minimum activity duration threshold of an activity. The *s_act* parameter roughly demonstrated the maximum distance range where an activity takes place. If the distance between two consecutive clusters stayed within *s_act*, it implies that these two clusters still belong to the same activity, and the location points fell within these two clusters would be labeled as activities, otherwise a trip will be generated. The *t_act* parameter defines the minimum duration for an activity. If the minimum time lag between two consecutive clusters is shorter than *t_act*, it implies no activity happens, which can

happen at traffic lights, traffic congestions, stop by, pick up, etc, otherwise an activity would be identified between the two clusters.

4.2.3 Non-Activity Adjustment

According to the field observations, the current technology enables the smartphone to automatically reduce the background activity when it is not active to save the battery, i.e. iPhone, One Plus. Therefore, when people go sleep at night, the sample rate would drop to a few minutes or hours until the person moves again. Therefore, an adjustment factor for the low sample rate at night, *t_gap*, was proposed. The *t_gap* checked the time gap between two consecutive location points within each trip.

4.3 Travel Mode Imputation with Machine Learning Algorithms

4.3.1 Overview of Machine Learning Algorithms

The objective of the travel mode imputation of this study is to identify drive, rail, bus, bike and walk travel modes. The air travel mode is not considered in this study, but it can be simply identified with heuristic rule using speed and time constraints. This study uses machine learning methods to impute travel modes with the feature generated from PCMDL data. Several machine learning methods will be examined in terms of prediction accuracy, including K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), eXtreme Gradient Boosting (XGB), Random Forest (RF), and Deep Neural Network (DNN).

4.3.1.1 K-Nearest Neighbors

K-Nearest Neighbors (KNN) is one of the earliest and simplest classification methods [75]. The main idea of KNN is to find top *k* nearest samples of target sample based on distance measurement. The distance between two samples is usually calculated based on Euclidean distance:

$$d(x_i,x_j)=\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+(x_{i3}-x_{j3})^2+\cdots+(x_{ip}-x_{jp})^2}$$

where x_i represents the sample with *p* features. Then the target sample will be classified as the majority class of these *k* neighbors. The key parameters in KNN are the *k* value and the distance measurement.

4.3.1.2 Support Vector Classifier

Support Vector Classifier (SVC) was developed by Cortes and Vapnik in the 1990s [76]. Numerous extensions of SVC were proposed and applied in the area of face recognition, pattern recognition, etc [77-79]. SVC can address the non-linearly separable samples by using the kernel function to map the data into a higher dimension, thus finding a hyperplane that best divides the data into different classes. Some examples of the kernel functions include polynomial, Gaussian, Gaussian radial basis function (RBF) and sigmoid, the equations of which are shown below:

$$\text{Polynomial: } K(x_i,x_j)=(x_i \cdot x_j+1)^p$$

$$\text{Gaussian: } K(x_i,x_j)=e^{\left(-\frac{\|x_i-x_j\|}{2\sigma^2}\right)}$$

$$\text{Gaussian radial basis: } K(x_i,x_j)=e^{(-\gamma\|x_i-x_j\|^2)},\ \gamma>0$$

Sigmoid: $K(x_i,x_j)=tanh(kx_i^T x_j+c),\ k>0$ and $c>0$

where x_i represents the i^{th} feature of the input features, p represents the degree of the polynomial, σ represents the standard deviation of the Gaussian distribution.

4.3.1.3 Extreme Gradient Boosting

Extreme Gradient Boosting (XGB) is one of the most recent ensemble-learning algorithms using the boosting technique [80,81]. The main idea of boosting is to train a set of weak classifiers using the same samples and then combine them into one strong classifier to improve the classification accuracy, where new classifiers are added to reduce errors based on previous models until no further improvements can be made [82,83].

4.3.1.4 Random Forest

Random Forest (RF) is one of the most famous ensemble-learning algorithms using the bagging technique [84,85]. Bagging (Bootstrap aggregating) is a machine learning technique that tends to improve the stability and accuracy of machine learning algorithms [85]. It generated multiple training sets by sampling from the data uniformly and with replacement. RF not only employs the bagging technique, but also used a modified tree learning algorithm that selects a random subset of the features without using all features, which is called feature bagging [84]. In short, RF is essentially a collection of decision trees [86] and each decision tree is trained with the different training sets and different features. The classification result follows the majority vote of all the decision trees in the forest.

4.3.1.5 Deep Neural Network

Deep Neural Network (DNN) is an Artificial Neural Network (ANN) with multiple layers between the input and output layers [87,88]. DNN can model the complex non-linear relationship between the input and the output by updating the weight vertices connecting each virtual neural between layers through back-propagation [89]. Detailed methodology of DNN and ANN can be found in [87,88]. Compared to an ANN, DNN can capture more complex non-linear relationships by adding more hidden layers. However, the neural network structure of a DNN needs to be designed efficiently and it also suffers from overfitting and computation time issues.

4.3.2 Feature Set Construction

Feature set construction directly affects the model performance. Before constructing the features, the air travel mode is filtered out using a rule-based method since it's easy to distinguish from the other travel modes. In this study, three thresholds are used to filter out the air trips from all the trips identified: average speed, trip time and trip distance. Here the 100 mph, 1 hour and 100 miles are selected as the value of these thresholds, indicating that for a trip, if the average speed exceeds 100 mph, the trip time is larger than 1 hour and the trip distance exceeds 100 miles, then this trip is considered as an air trip.

In this study, the features are constructed using the information derived from each trip and impute the travel mode using the trip-based approach, including three categories of features as shown in Table 4-1: sample rate feature, trip feature, and multimodal transportation network features. Though more features can be incorporated into the

feature set, it might not be computationally expensive to implement on the large-scale LBS dataset.

The main idea of having the sample rate feature is that most LBS data's sample rate is random since it is generated when smartphone users use the location-based services. Here the average number of records per minute is used to represent this randomness.

The trip features describe the characteristics including the trip distance, origin-destination distance, trip time, average speed, minimum speed, maximum speed, median speed, and 5/25/75/95 percentile speed. Though acceleration related information is proved to be effective in literature [61-67], it is not considered in this study. The main reason is that the acceleration can be calculated with high-frequency data, while the frequency PCMDL data, LBS data, in particular, ranges from one second to minutes or even hours. Under this consideration, it might result in biased estimation in terms of acceleration.

The multimodal transportation network features are also considered as the import features to distinguish between different travel modes [52,62]. Here, a 10-meter buffer is generated using the multimodal transportation networks (bus, rail and drive) in order to obtain the percentage of records for each trip that fell within the networks respectively. To enhance the bus mode imputation results, a 50-meter buffer for the bus stops is also generated to check the percentage of records for each trip that fell within the bus stops.

Table 4-1. Features used in Travel Mode Imputation.

Features	Unit
Sample Rate Feature	
Average # of records per minutes	number / minute
Trip Features	
Trip distance	meters
Origin-Destination distance	meters
Trip time	minutes
Average speed	meters
Minimum speed	meters
Maximum speed	meters
Median speed	meters
5-percentile speed	meters
25-percentile speed	meters
75-percentile speed	meters
95-percentile speed	meters
Multimodal Transportation Network Features	
% of records fell within 10 meters of the rail network	percentage
% of records fell within 10 meters of the bus network	percentage
% of records fell within 10 meters of drive network	percentage
% of records fell within 50 meters of bus stops	percentage

4.3.3 Synthetic Minority Over-sampling Technique

In order to balance the travel mode classes in the training dataset, the Synthetic Minority Over-sampling Technique (SMOTE) is used to address the imbalanced dataset via synthesizing the minority class from the existing samples [90].

Figure 4-4 illustrates the basic idea of SMOTE. The synthetic sample is generated with the following steps: first, for each sample in each minority class, the distance to all the other sample in the same minority class is calculated in order to obtain its k nearest neighbors; then, a sample multiplier N is defined based on the imbalance ratio. For each minority sample x, randomly select samples from its k nearest neighbors x_n; finally, generate the new sample with the equation as follows:

$$x_{new} = x + rand(0,1) - |x - x_n|$$

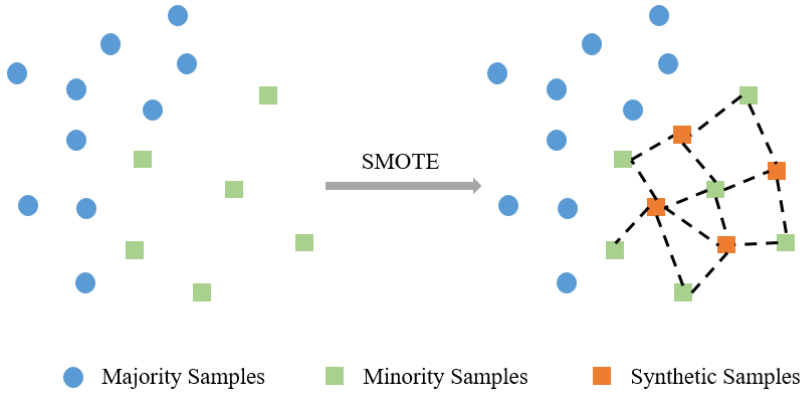


Figure 4-4. Illustration of SMOTE.

4.4 Model Relaxation

The aforementioned methods will be calibrated and compared using the two incenTrip datasets as shown in the following section. However, as mentioned in the literature, the sample rate is different between GPS data with user recall and LBS data. Therefore, the method developed based on the incenTrip data cannot be directly applied to the LBS data. Here this study relaxes the model constraints in order to enable the model to be applicable to the LBS data.

For the trip end identification part, two of the six parameters are changed in order to take the low sample rate problem of LBS into account. The temporal threshold t will be increased and the minimum neighbor's m will be decreased to capture more clusters in the LBS data.

For the travel mode imputation part, the multimodal transportation network features are relaxed, where the 10-meter network buffer is expanded to 50-meter network buffer to cover the sparse LBS data.

Chapter 5: Results

This chapter shows the model development and model application results with the proposed framework. Firstly, the model development part uses the first incenTrip application dataset to validate the trip end identification. Then, the second incenTrip dataset is used to train the travel mode imputation methods with machine learning methods. The performance of the models is examined. Then the proposed framework is relaxed and applied to two LBS datasets to validate against travel surveys. The trip distance and trip time distribution are used as criteria to validate the trip end identification results. The mode share at different geographic levels are used for travel mode imputation results comparison.

5.1 Model Development using incenTrip Application Data

5.1.1 Trip End Identification Parameter Calibration and Result

As introduced in the previous section, the proposed framework is capable of different sample rates. In the case study, the trace segmentation purpose was set to include short activities, such as waiting for transferring at a metro station, waiting for the bus at a bus stop etc..

To satisfy the trace segmentation purpose and at the same time conform to each sample rate, five parameters were calibrated, including the spatial threshold *s*, temporal threshold *t*, minimum neighbors *n*, maximum distance threshold for an activity *s_act* and minimum duration of an activity *t_act*. The adjustment factor *t_gap* was fixed at 300 s since it's only used for very few irregular data. *s* determines the distance range

of a stop, thus increasing the value of *s* would yield more identified stop since more location points would be involved in the clusters. To ensure all the stops were captured including traffic congestion and waiting at the traffic light for both vehicle and pedestrian, four constraints were added as shown below:

$$t \geq n * f_1$$

$$t\_act \geq n * f_2$$

$$s\_act \geq s$$

$$n * f_2 \geq s / v$$

where v is the average walking speed, here we consider 1 m/s; f_1 is sample rate when not moving; f_2 is the sample rate when moving. Consider the real-world scenario when a person stops, it is intuitively to set the *s* value to be relatively small. With domain knowledge, here we use 25-meter, 50-meter and 100-meter as the candidate *s* value. Also, the *t_act* was set as 300 seconds to obtain most of the short activities. Then, with the given sample rate, the corresponding range for other parameters could be calculated. For each sample rate, two testers were selected to calibrate the parameters. The two testers have different daily mobility patterns, with one typical driver and the other one a typical public transportation user. Different combinations of the parameters were applied to these two testers to measuring the difference between reported trips and identified trips.

Table 5-1 shows the calibrated parameters used in the case study for each sample rate. The proposed framework was further applied to all data collected during the testing period. Similar to the conclusion in [8,9], even the testers were required to record their trip diary correctly, highly-biased trip start time and trip end time and under-reported

trips problems were observed frequently when compared to each tester's trip diary. To reduce the noise and validate the proposed framework, therefore, three people's data with the best data consistency and recorded travel time was selected by comparing the trip diaries and identified trips with the visual check on the GIS platform. Then, testers were asked to re-confirm both their reported trips and underreported trips identified from the location data. It should be noted that for the 1s/1s sample rate, only two person's data were available due to the short testing period, thus only 23 reported trips are included here.

Table 5-1. Calibrated Parameters for Each Sample Rate.

Sample Rate	*s* (m)	*t* (s)	*m*	*s_act* (m)	*t_act* (s)	*t_gap* (s)
1s/1s	50	100	50	100	300	300
2s/5s	50	200	25	100	300	300
5s/10s	50	200	15	100	300	300
5s/30s	50	500	15	100	300	300
15s/30s	50	600	10	100	300	300

Figure 5-1 and Table 5-2 show the trip end identification result. For each sample rate, the proposed framework was able to capture over 90% of the reported trips, with the overall hit-ratio for all sample rate is 94.5%. In addition, the proposed framework was also able to identify the underreported trips from the raw data. It can be observed that for each sample rate, about 15% to 35% of the trips were not reported by the respondents. Identifying these underreported trips help produce a more detailed mobility pattern and trip chain of each respondent.

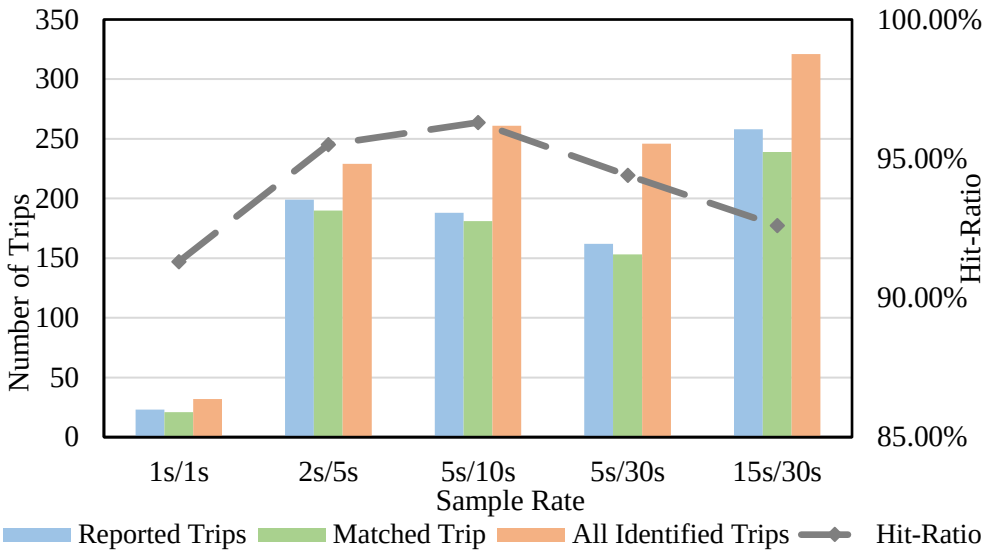


Figure 5-1. Trip End Identification Result.

Table 5-2. Trip End Identification Result.

Sample rate	Reported Trips	Matched Trip	All Identified Trips	Hit-Ratio
1s/1s	23	21	32	91.30%
2s/5s	199	190	229	95.50%
5s/10s	188	181	261	96.30%
5s/30s	162	153	246	94.40%
15s/30s	258	239	321	92.60%
Total	830	784	1089	94.50%

5.1.2 Travel Mode Imputation Result

The dataset two mentioned in section 3.1 is used to train the travel mode Imputation using KNN, SVC, XGB, RF and DNN respectively. 70% of the data is used for training and 30% of the data is used for testing. The SMOTE is then applied to the training data in order to address the imbalanced sample problem. For each machine learning method, a parameter grid is created to conduct a random search among different parameter combinations to fine-tune the models. Table 5-3 shows the parameters tuned in this study. For DNN, a simple structure is used to train the model.

Table 5-3. Parameter Grid for Machine Learning Models

Model	Parameter Grid
KNN	k: [1,5,10,20,30,40,50] weights: [‘uniform’,’distance’] algorithm: [‘ball_tree’,’kd_tree’,’brute’] leaf_size: [1,10,30,50,100] p: [1,2,3,4]
SVC	Cs: [0.001,0.01,0.1,1,10,100] gammas: [0.001,0.01,0.1,1] class_weight: [‘None’,’balanced’]
XGB	colsample_bytree: U(0.7,0.3) gamma: U(0,0.5) learning_rate: U(0.03,0.3) max_depth: [2,3,4,5,6] n_estimators: [100-150] subsample: U(0.6,0.4)
RF	n_estimators: [100-1000] max_features: [‘auto’,’sqrt’] max_depth: [10,110] min_samples_split: [2,5,10,15,20] min_samples_leaf: [1,2,4,6,8] bootstrap: [‘True’,’False’] class_weight: [‘None’,’balanced’,’balanced_subsample’]
DNN	activation function: [‘ReLU’]

** U represents uniform distribution.*

During the model training process, 10-fold cross-comparison is used to evaluate the model performance. In this study we used the F1 score to evaluate the model performance, which is calculated as shown below:

$$F1=2\cdot\frac{Precision\cdot Recall}{Precision+Recall}$$

$$Precision=\frac{TP}{TP+FP}$$

$$Recall=\frac{TP}{TP+FN}$$

where TP represents the true positive, FP represents false positive, and FN represents false negative.

Figure 5-2, Figure 5-3 and Table 5-3 compare model performances using the F1 score. Among all the tested machine learning methods, RF achieved the highest F1 score across all travel modes in both four and five modes models. The bus mode has the least prediction accuracy among the modes, mainly because the drive trips and bus trips are similar to each other. Detailed confusion matrixes are listed in Appendix A.

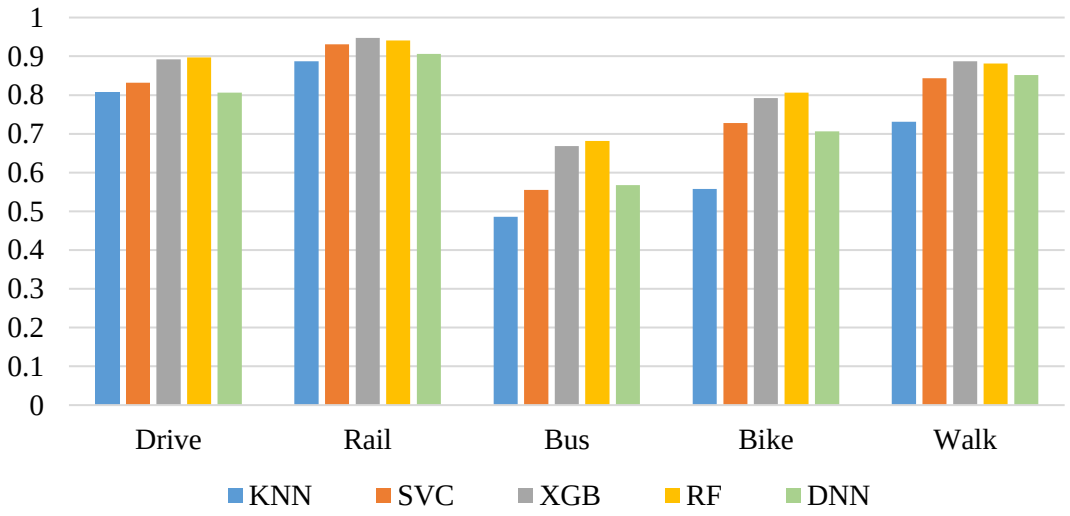


Figure 5-2. Travel Mode Imputation Result with Five Travel Modes.

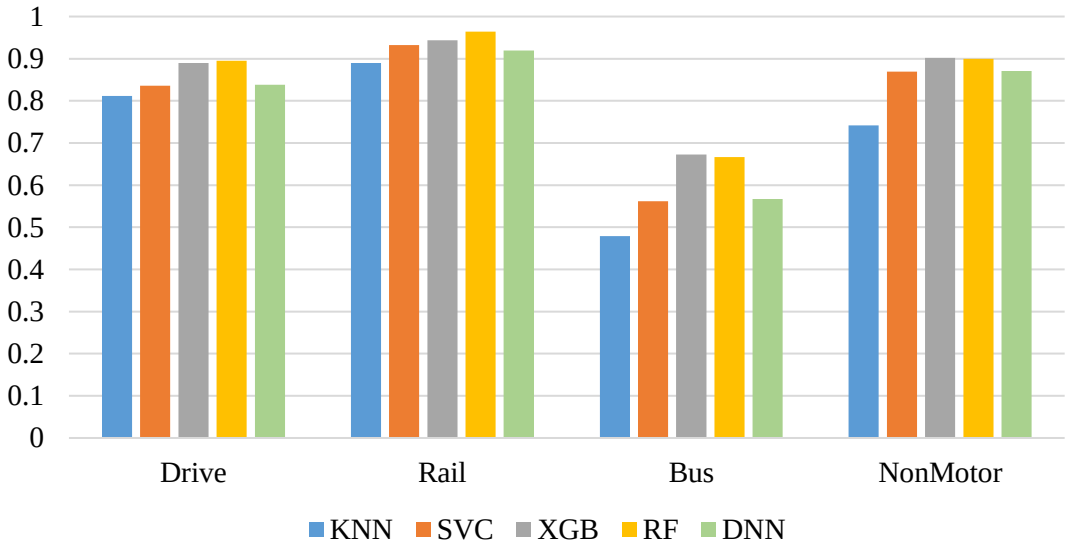


Figure 5-3. Travel Mode Imputation Result with Four Travel Modes.

Table 5-4. Model Performance Comparison (F1 Score)

		KNN	SVC	XGB	**RF**	DNN
Four Modes	Drive	0.81	0.83	0.89	**0.90**	0.81
	Rail	0.89	0.93	0.95	**0.95**	0.91
	Bus	0.49	0.55	0.67	**0.68**	0.57
	Bike	0.56	0.73	0.79	**0.81**	0.71
	Walk	0.73	0.84	0.89	**0.89**	0.85
Five Modes	Drive	0.81	0.84	0.89	**0.90**	0.84
	Rail	0.89	0.93	0.94	**0.96**	0.92
	Bus	0.48	0.56	0.67	**0.67**	0.57
	NonMotor	0.74	0.87	0.90	**0.90**	0.87

Figure 5-4 and Figure 5-5 shows the feature importance value of the RF model for five and four travel modes respectively. The feature importance value (Gini importance) is automatically calculated using the python package sklearn, representing each importance as the sum over the number of splits across all trees. It can be seen that the speed variables (95 quantile speed, maximum speed and average speed) are the most important. Also, the percentage of records which fell within 10 meters of the rail network is also significantly important since it is a representative feature for imputing rail trips.

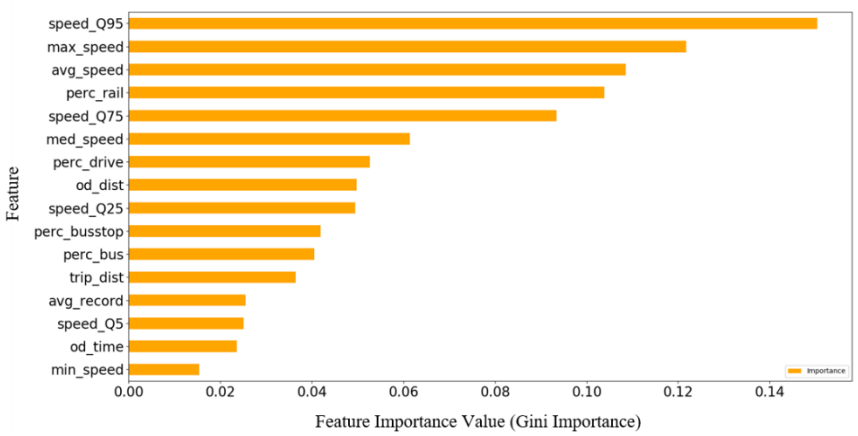


Figure 5-4. RF Feature Importance Value for Five Travel Modes.

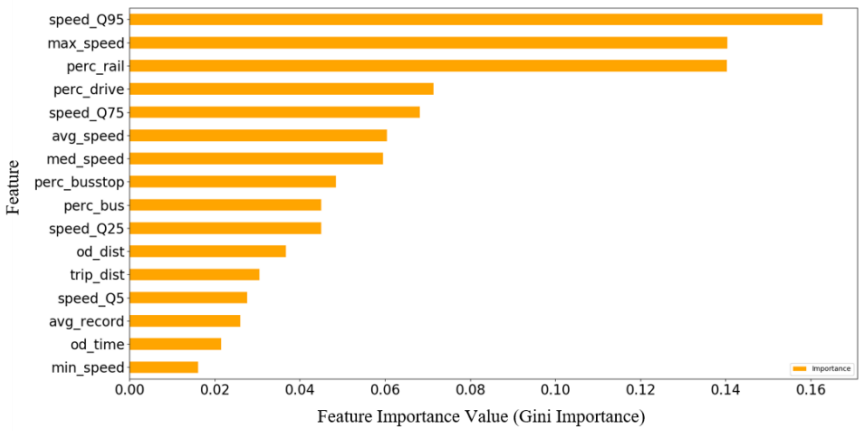


Figure 5-5. RF Feature Importance Value for Four Travel Modes

5.2 Case Study One: Application on Maryland Location-based Service Data Sample

In case study one, the models developed from the incenTrip application data are relaxed as described in section 4.4 and applied on the Maryland LBS dataset. Based on the 15s/30s sample rate from the incenTrip data, three model relaxations are made: (1) the temporal threshold is relaxed from 600 seconds to 1800 seconds; (2) the minimum neighbors is relaxed from 10 to 5; (3) the three multimodal transportation network buffer (drive, rail, and bus) is relaxed from 10 meters to 50 meters. The trip distance and trip time distribution are compared against the 2017 NHTS results in for all trips originated from Maryland and Washington D.C. The mode share is first compared at a statewide level using the 2017 NHTS mode share in Maryland and Washington D.C. Then it is also compared against the 2007/08 BMC-TPB HHTS at the county level. The NHTS mode share results at the county-level is also used as a supplement comparison source for mode share. A visual comparison is provided at the census tract level.

5.2.1 Trip Distance and Trip Time Distribution Comparison

According to USDOT BTS [4], the trips can be divided into short-distance trips and long-distance trips using the 50 miles threshold. Thus, the trip distance distribution for the LBS data is compared for both short-distance (<50 miles) and long-distance (≥50 miles) trips. It should be noted that the 2017 NHTS uses Google API to calculate the distance for each trip, which underestimates the trip distance [1]. At the same time, trips extracted from the LBS data use the great circle distance accumulated from consecutive location points within each trip, thus also underestimating the trip distance.

Figure 5-6 and Figure 5-7 show the short-distance and long-distance distribution comparison between NHTS results and LBS data results respectively. For both short-distance and long-distance trips. For short-distance trips, the trip distance distribution is similar between NHTS results and LBS data results. For long-distance trips, the LBS result shows more long-distance trips than the survey.

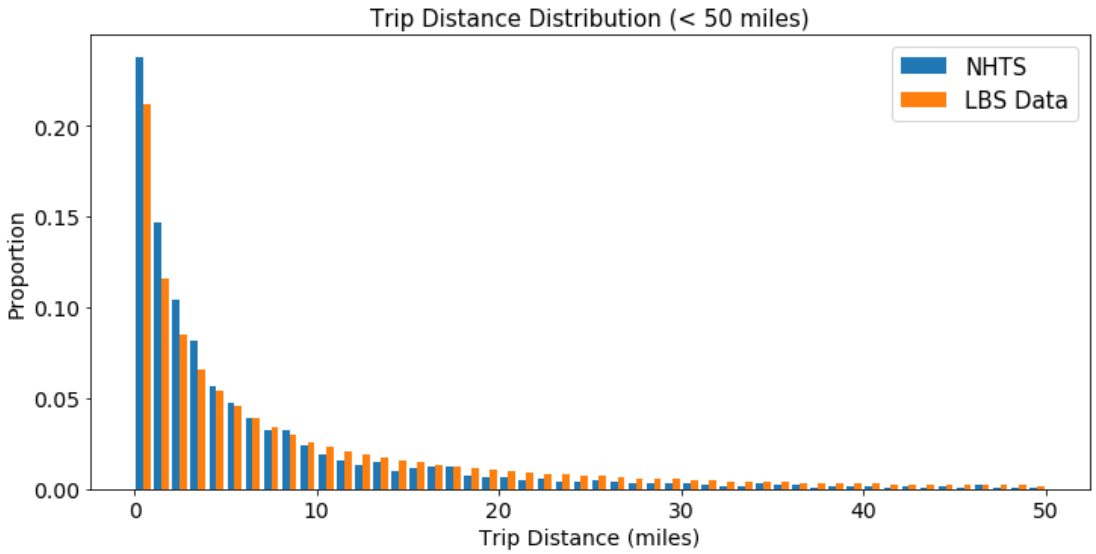


Figure 5-6. Maryland Trip Distance Distribution for Short-Distance Trips.

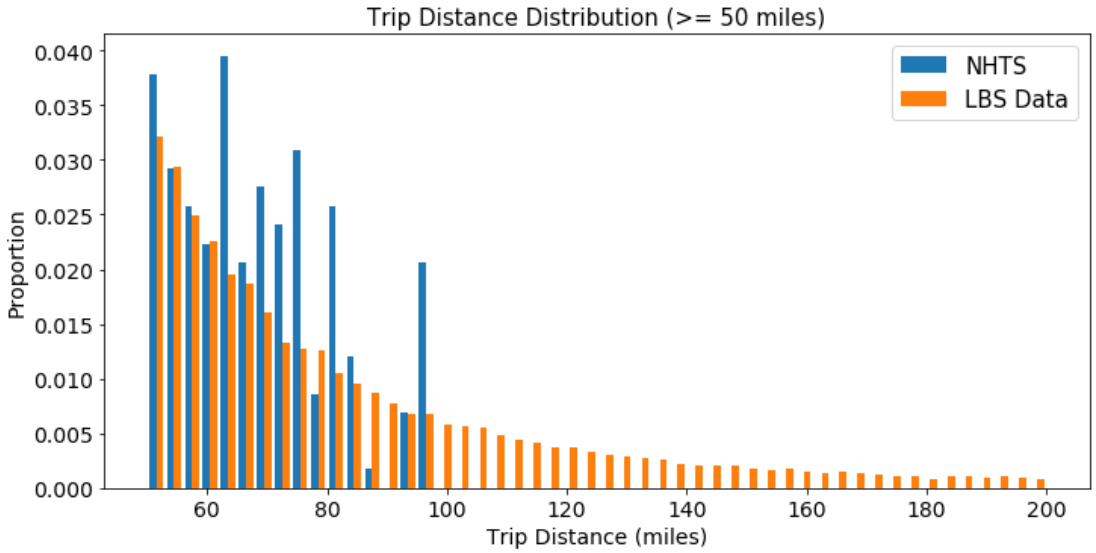


Figure 5-7. Maryland Trip Distance Distribution for Long-Distance Trips.

Figure 5-8 shows the trip time distribution comparison. The overall trend is similar, while the LBS data underestimates trips around 10 minutes and overestimates long trips. This can mainly be attributed to the stochastic sample rate problem of LBS data.

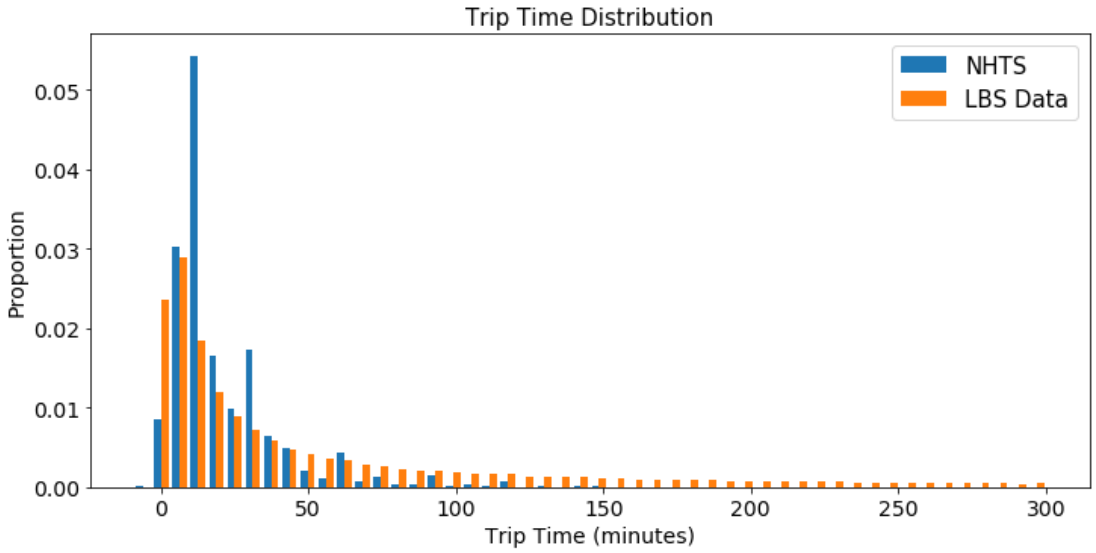


Figure 5-8. Maryland Trip Time Distribution.

5.2.2 Statewide Mode Share Comparison

Figure 5-9 shows the statewide mode share comparison result. The overall mode share distribution is also consistent with the 2017 NHTS mode share. The drive and rail mode shares are perfectly matched with the 2017 NHTS. The bus mode share estimated from the LBS data is relatively low, which might also be the reason for the incomplete bus network. The non-motorized mode share is slightly higher.

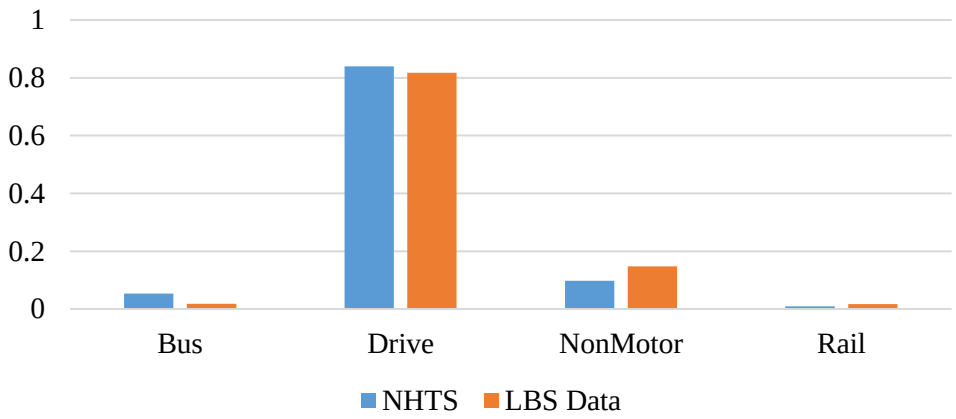


Figure 5-9. Statewide Mode Shares.

5.2.3 CBSA-Level Mode Share Comparison

The CBSA-level mode share is also compared against the 2017 NHTS more shares. Two CBSAs fall within Maryland and Washington D.C. are examined: Baltimore Metropolitan Council (BMC) and Washington Metropolitan Area (DMV). The trips originated from these two CBSAs are extracted and to obtain the mode shares from the 2017 NHTS.

Figure 5-10 illustrates the comparison results. The overall mode shares estimated from the LBS data for these two CBSAs show similar trends compared to 2017 NHTS. For BMC, the drive and rail mode share match perfectly, while the bus travel is

underestimated and the non-motorized travel is overestimated. For DMV, both rail and bus trips are underestimated.

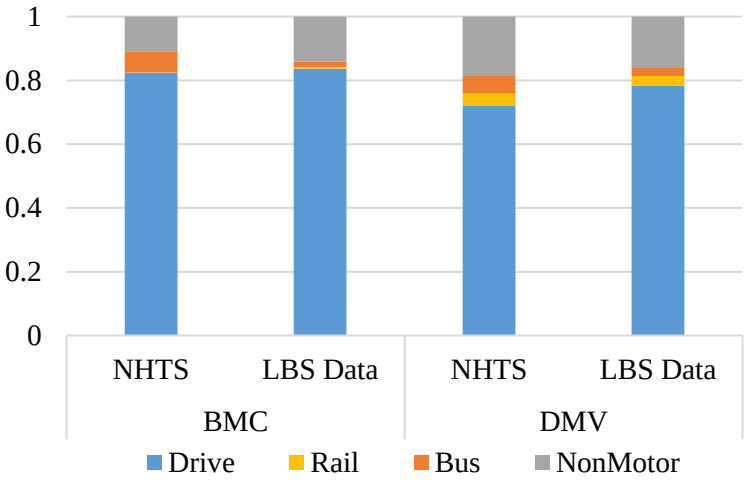


Figure 5-10. CBSA-Level Mode Shares

5.2.4 County-Level Mode Share Comparison

The mode shares are further compared at the county-level with two travel surveys. The first travel survey is the 2007/08 TPB-BMC HHTS that covers Anne Arundel County, Baltimore County, Baltimore City, Carroll County, Harford County, Howard County, Montgomery County, Prince George's County and Washington D.C. Figure 5-11 shows the comparison results. For Washington D.C., the rail mode share estimated from the LBS data is lower than the survey and the others are matched perfectly. For the other eight counties, the bus mode share estimated from LBS data is lower, and the non-motorized mode share is higher. Figure 5-12 shows the correlation between the estimated mode shares and the survey mode shares. It can be seen that there is a high correlation observed between our estimates and ground truth. However, it should be noted that the 2007/08 TPB-BMC HHTS was conducted over ten years ago and the travel patterns might have changed greatly.

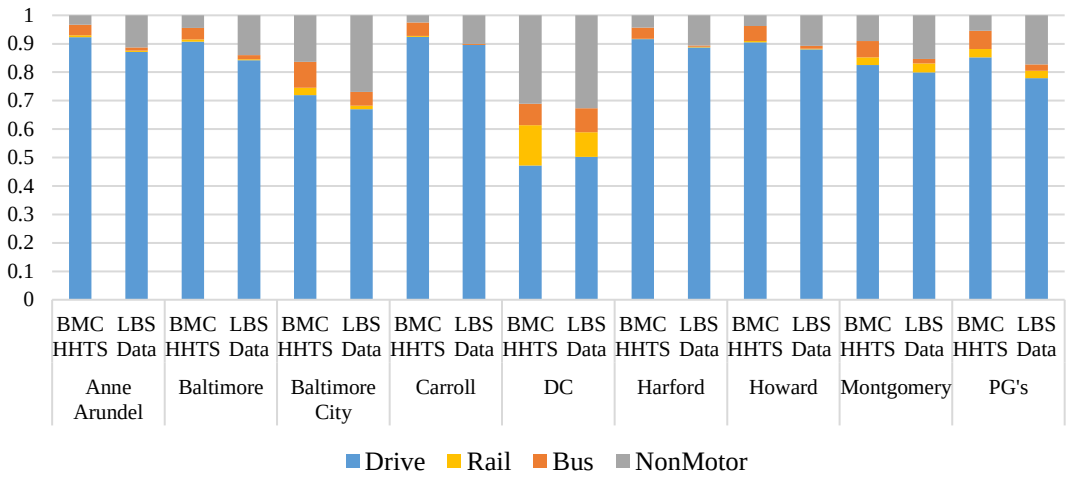


Figure 5-11. 2007/2008 TPB-BMC HTS County-Level Mode Shares.

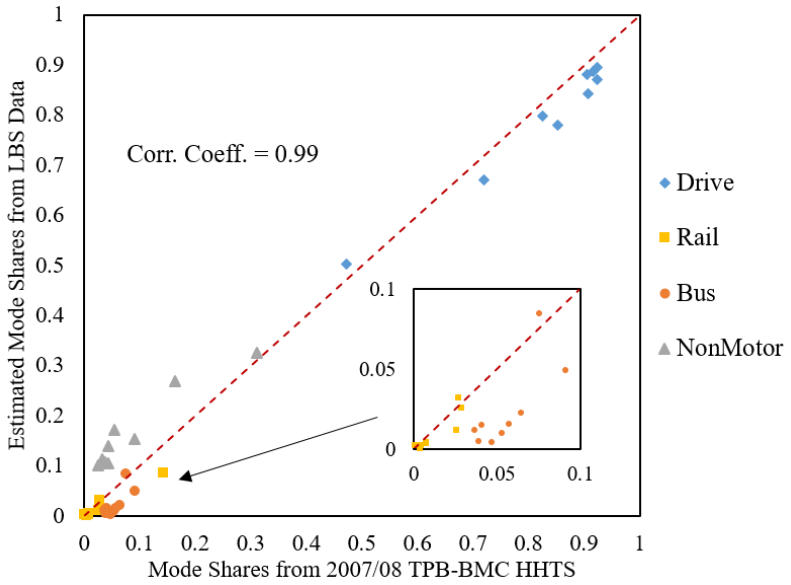


Figure 5-12. Correlation between Estimated Mode Shares and 2007/08 TPB-BMC HHTS Mode Shares.

The estimated mode shares are also compared to the 2017 NHTS mode shares using the trips originated from every county in Maryland and Washington D.C. The 2017 NHTS mode shares might better reflect the recent travel patterns while for some counties it might suffer from the biased estimation because of small samples. Figure 5-13, Figure 5-14 and Figure 5-15 show the comparison results. It can be observed that for some counties, the NHTS mode shares are biased which cannot be used for

comparison, such as Charles County and Kent County. The bus travel is underestimated over most areas except for Washington D.C. For Washington D.C., the bus and rail mode shares match perfectly. The high drive mode share might because lots of drive trips originated from Washington D.C. and ended out of Washington D.C. are observed that might not be taken into account in the survey.

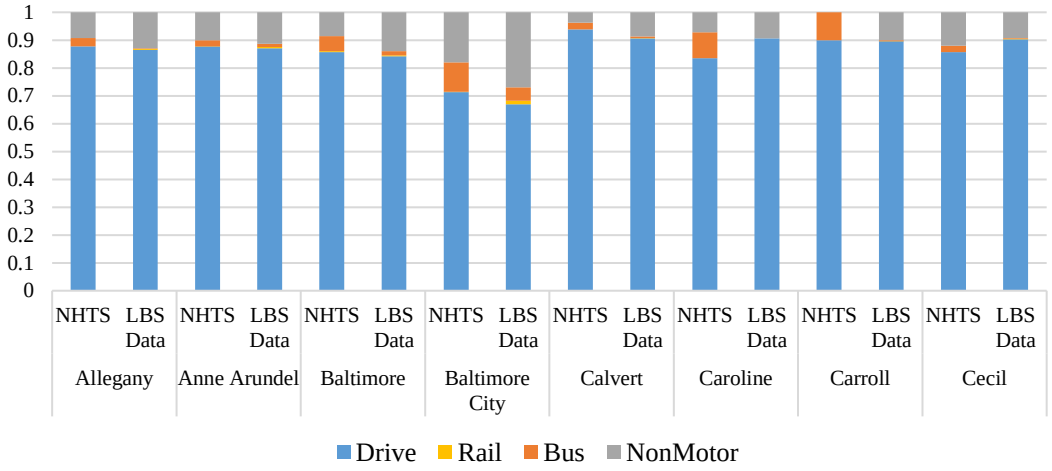


Figure 5-13. NHTS County-Level Mode Shares (1).

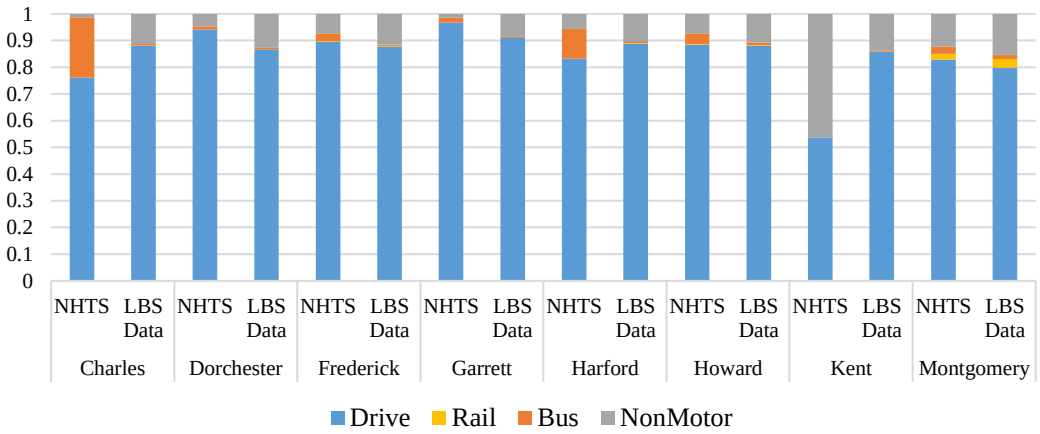


Figure 5-14. NHTS County-Level Mode Shares (2).

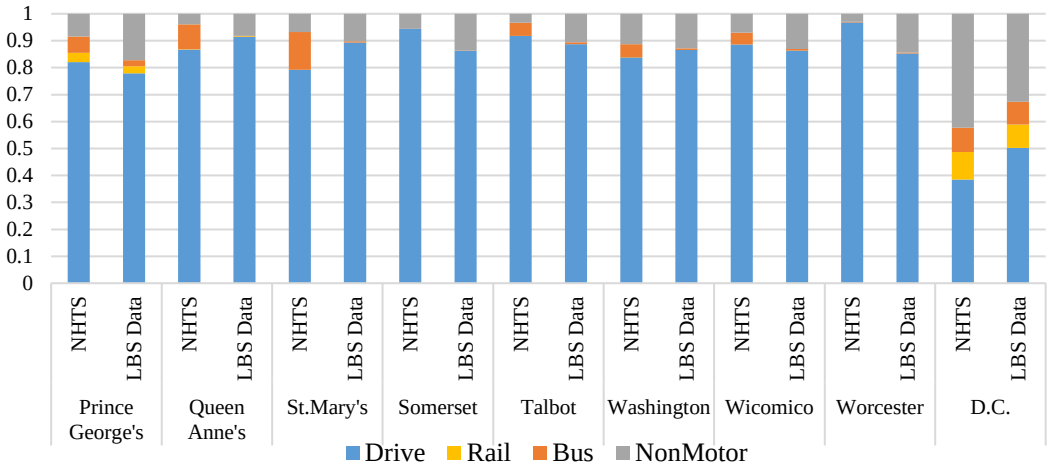
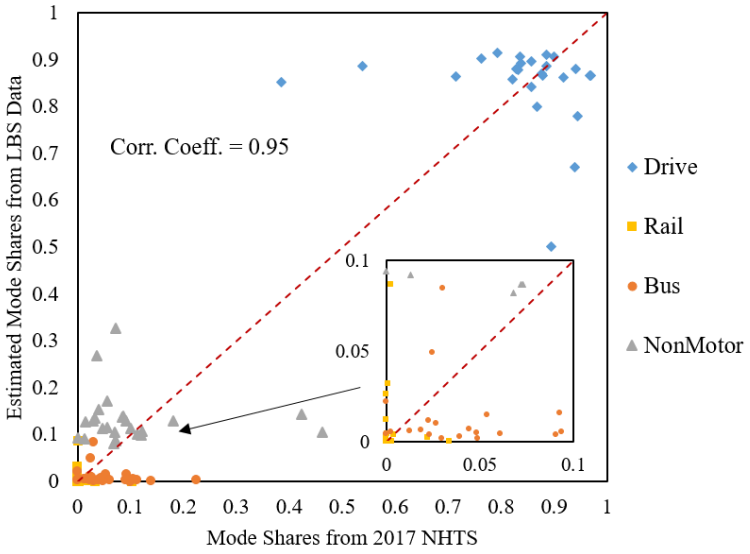


Figure 5-15. NHTS County-Level Mode Shares (3).

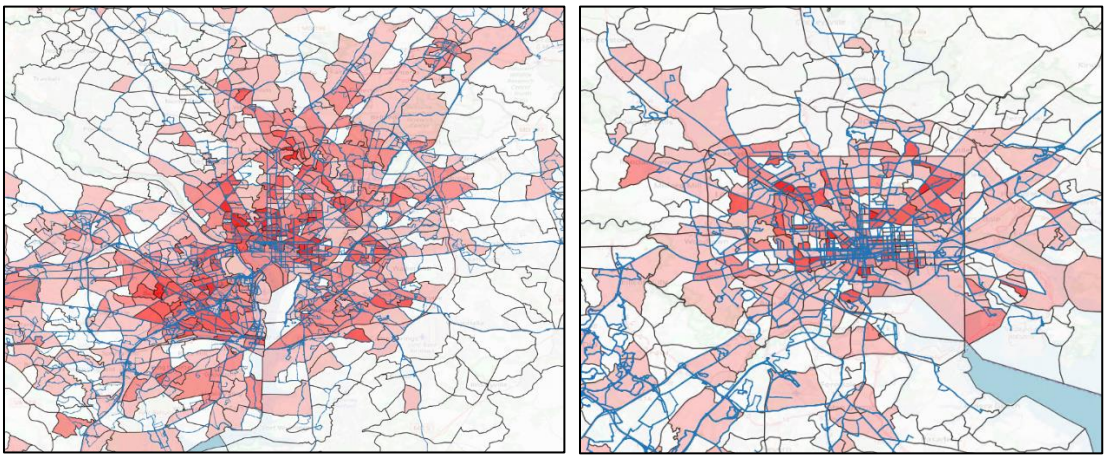


5.2.5 Census Tract-Level Mode Share Comparison

Since the estimated mode share at census tract-level is not available from any data sources, the census tract-level rail and bus mode shares estimated from the LBS data are also plotted for illustration purposes, as shown in Figure 5-16 and Figure 5-17. The figures are plotted using the Jenks natural breaks optimization [91] and the depth of color represents the magnitude of the value regarding mode shares.

(a) (b)

Figure 5-16. Census Tract-Level Rail Mode Shares: (a) Washington D.C.; (b) Baltimore City.



(a) (b)

Figure 5-17. Census Tract-Level Bus Mode Share Comparison. (a) Washington D.C.; (b) Baltimore City.

For both Washington D.C. and Baltimore city, the mode share distribution of census tracts is highly correlated with the geographical distribution of rail and bus networks. In other words, the values of rail and bus mode shares of census tracts highly depend on closeness to the rail and bus network. Also, since Washington D.C. has denser rail and bus networks, the relative mode shares are higher than Baltimore City.

5.3 Case Study Two: Application on the United States National Location-based Service Data Sample

In case study two, the exactly same models with relaxations used in case study one is applied to the National LBS dataset. The trip distance and trip time distribution are compared against the 2017 NHTS results. The mode share is also compared against the 2017 NHTS mode shares at a nationwide and state level. A visual comparison is provided at the Core-based Statistical Area (CBSA) level.

5.3.1 Trip Distance and Trip Time Distribution Comparison

Figure 5-17 and Figure 5-18 show the trip distance distribution comparison between NHTS result and LBS data result for short-distance and long-distance trips, respectively. For both short-distance and long-distance trips, the overall trip distance distribution is similar between NHTS results and LBS data results. For short-distance trips, trips shorter than 5 miles are underestimated and trips longer than 10 miles are overestimated. For long-distance trips, trips under 100 miles are underestimated and the others are overestimated.

Figure 5-20 shows the trip time distribution comparison. The overall trend is similar. The short trips are underestimated and the long trips are overestimated. The main reason for this result is because the stochastic sample rate of the LBS data which produces inaccurate trip time.

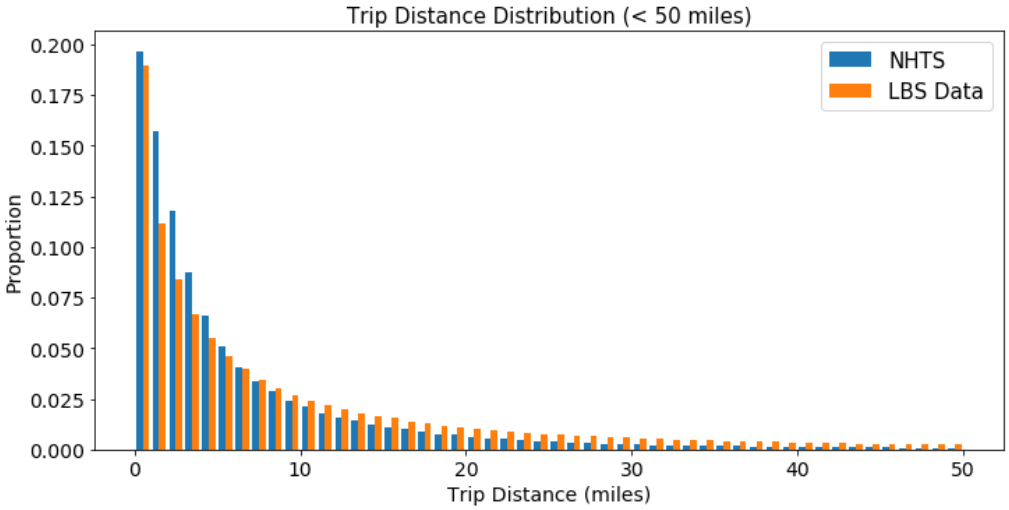


Figure 5-18. National Trip Distance Distribution for Short-Distance Trips.

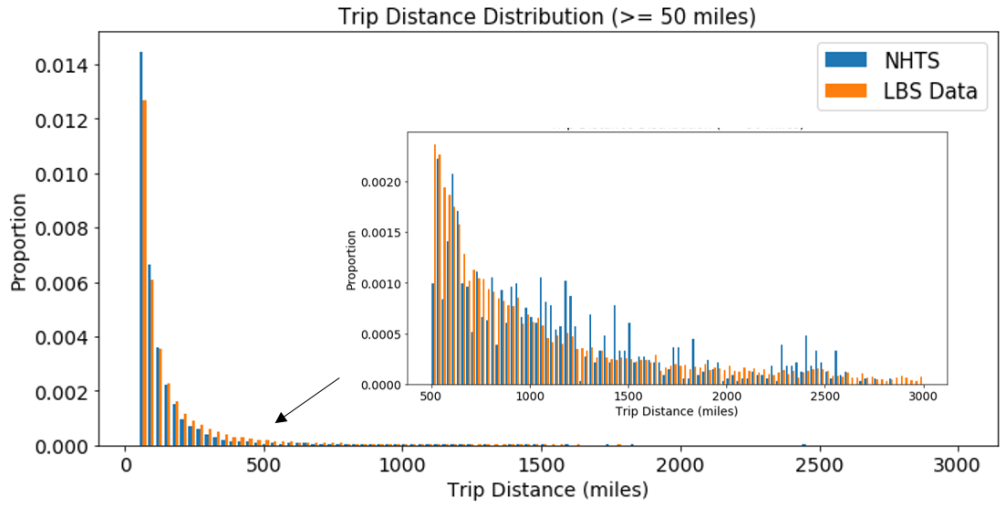


Figure 5-19. National Trip Distance Distribution for Long-Distance Trips.

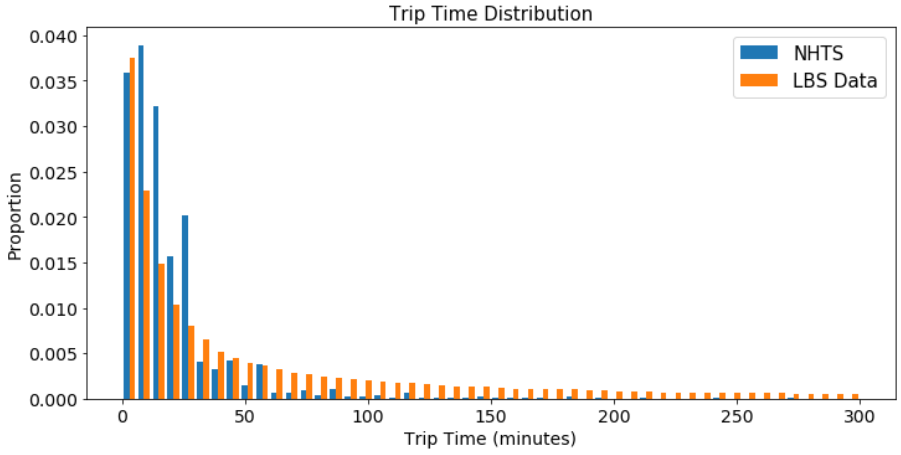


Figure 5-20. National Trip Time Distribution.

However, since biases also exist in NHTS, the trip distance and trip time distribution results from the LBS data are not supposed to perfectly match with the NHTS results. The overall trends should be similar while differences are acceptable since the population level ground truth information is not available.

5.3.2 Nationwide Mode Share Comparison

The air trips are firstly filtered out using the rule-based method as mentioned in the previous chapter. The result is compared to the top airport ranked by passengers boarded summarized by USDOT BTS [4], the top 10 of which are shown in Table 5-5. Figure 5-21 shows the heat map of all identified air trip origins, where the depth of the color represents the number of air trips originated from the closest airport. It can be observed that all the major airports are captured.

Table 5-5. Top 10 U.S. Airport Ranked by Passengers Boarded.

Rank	Airport	City	Rank	Airport	City
1	ATL	Atlanta	6	JFK	New York
2	LAX	Los Angeles	7	SFO	San Francisco
3	ORD	Chicago	8	SEA	Seattle
4	DFW	Dallas	9	LAS	Las Vegas
5	DEN	Denver	10	MCO	Orlando

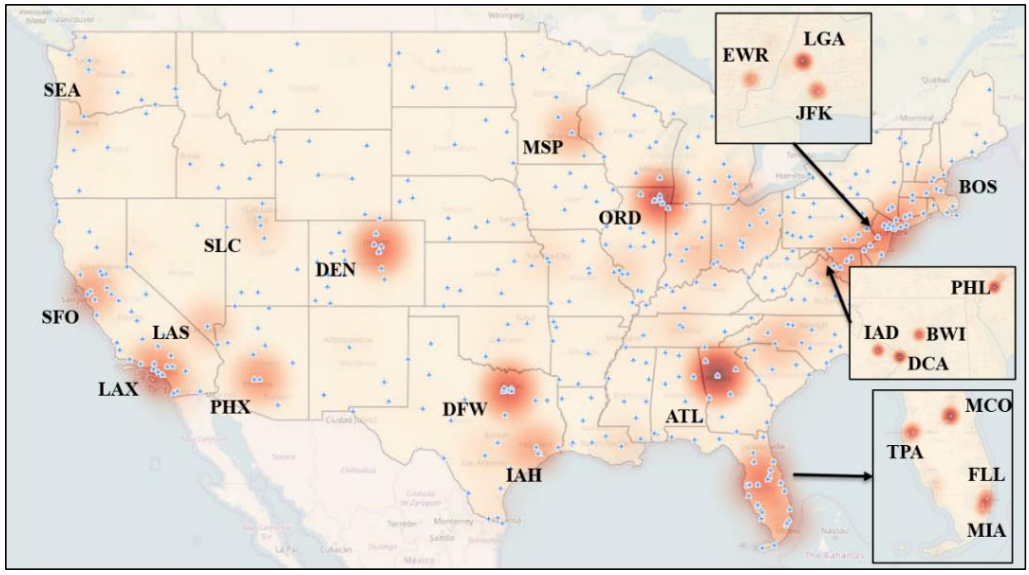


Figure 5-21. Nationwide Air Trips by Origins Heat Map.

Figure 5-22 shows the nationwide mode share comparison results. The overall mode share distribution is consistent with the 2017 NHTS mode share. The bus mode share estimated from the LBS data is relatively low, which might be the reason for the incomplete bus network. The non-motorized mode share is higher.

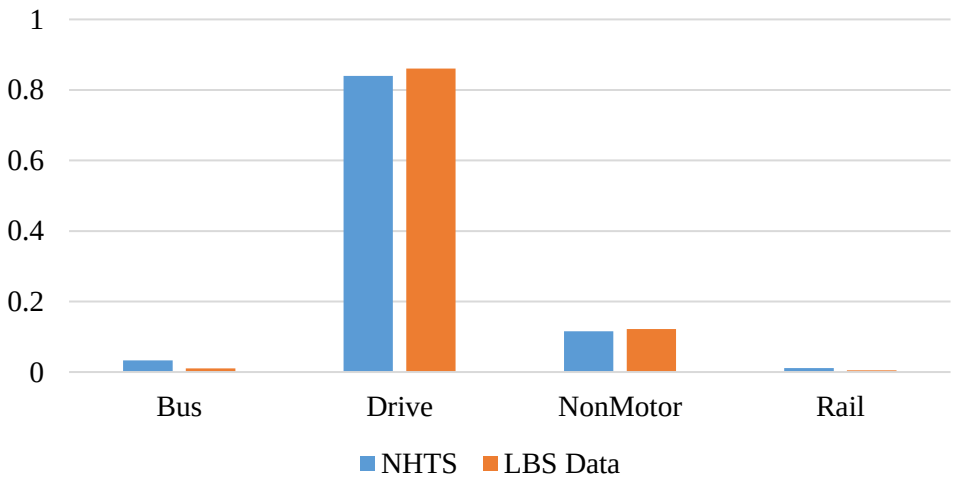


Figure 5-22. Nationwide Mode Shares.

5.3.3 State-Level Mode Share Comparison

Figure 5-23, Figure 5-24 and Figure 5-25 show the state-level mode share comparison for 50 states and Washington D.C. in the United States. The overall mode share across the U.S. is reasonable, with a slight underestimation of bus travel and overestimation of the non-motorized travel. The underestimation of bus travel might due to the incomplete national bus network. For non-motorized travel, since respondents tend to underreport short trips [8,9], which are most likely to be short walking and biking trips that can be detected from the LBS data, the result from LBS might reflect the real world more precisely.

In addition, the mode share estimated in Washington D.C. perfectly matches the survey compared to other states with relatively bus and rail mode share (IL, MA and NY). This might because the travel mode imputation model is trained using the data collected in the same region. Figure 5-26 shows the correlation between the estimated mode shares and the survey mode shares. It can be seen that there is a high correlation observed between our estimates and ground truth.

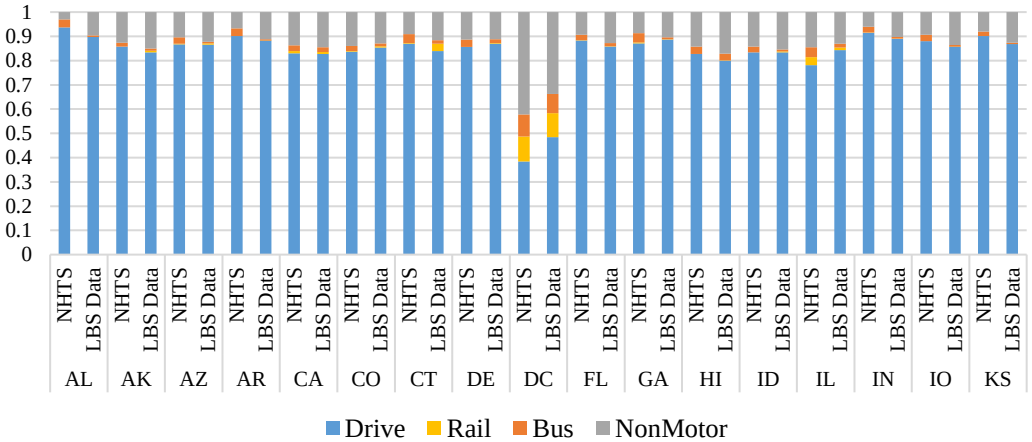


Figure 5-23. State-Level Mode Shares (1).

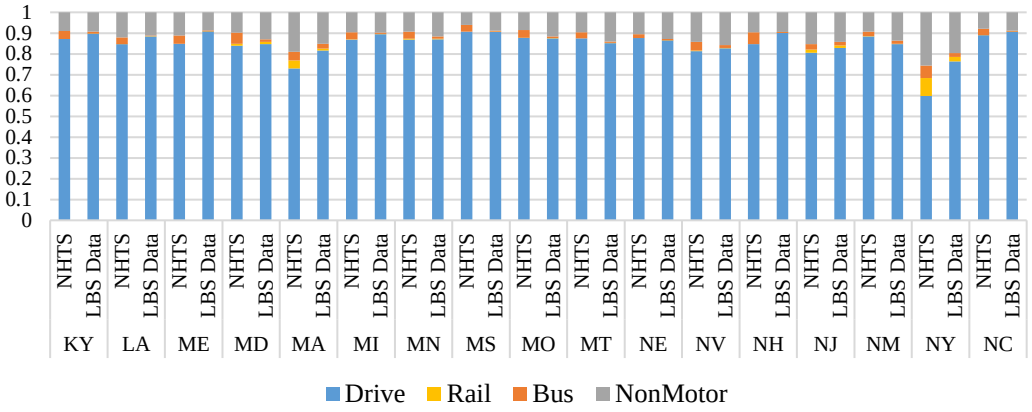


Figure 5-24. State-Level Mode Shares (2).

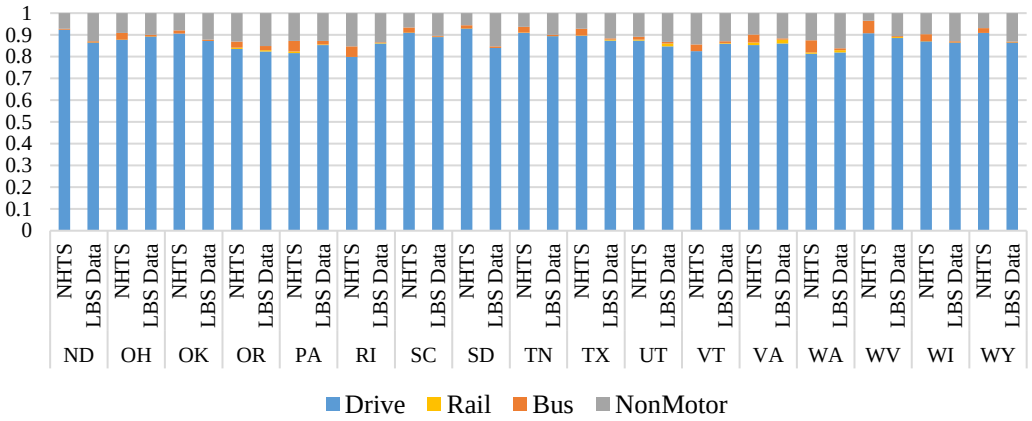


Figure 5-25. State-Level Mode Shares (3).

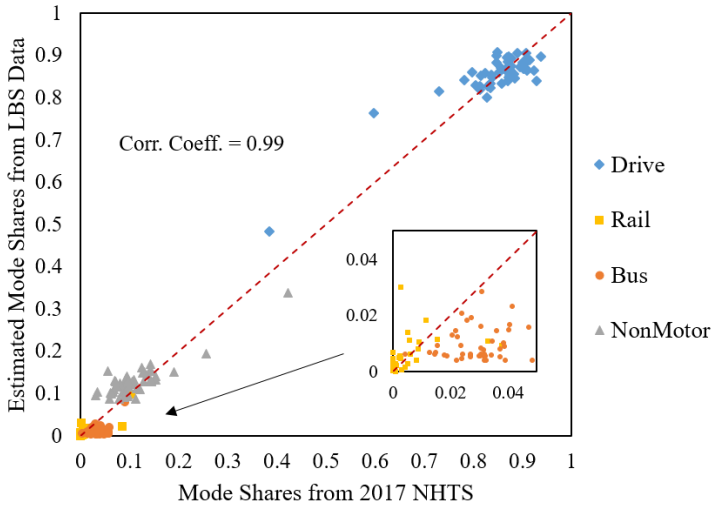


Figure 5-26. Correlation between Estimated Mode Shares and 2017 NHTS Mode Shares

5.3.4 CBSA-Level Mode Share Comparison

Since the estimated mode shares at CBSA level are biased due to limited sample, the CBSA-Level rail and bus mode shares estimated from the LBS data are also plotted for illustration purposes, as shown in Figure 5-27 and Figure 5-28. The figures are also plotted using the Jenks natural breaks optimization [91], where each class-s average deviation from the class mean is minimized and each class's deviation from the means of the other groups is maximized. The depth of color represents the magnitude of the value regarding mode shares.

Since the travel mode imputation algorithm is developed based on multimodal transportation networks, the imputation results for rail and bus travel modes are highly rely on the density of the rail and bus networks. For rail mode share, it can be observed that some typical CBSAs with well-developed rail or metro networks, such as Washington D.C., New York, Boston, San Francisco and Portland have obvious higher rail mode shares than the other CBSAs. For bus mode share, a similar trend is observed too where CBSAs with denser bus networks have higher bus mode shares.

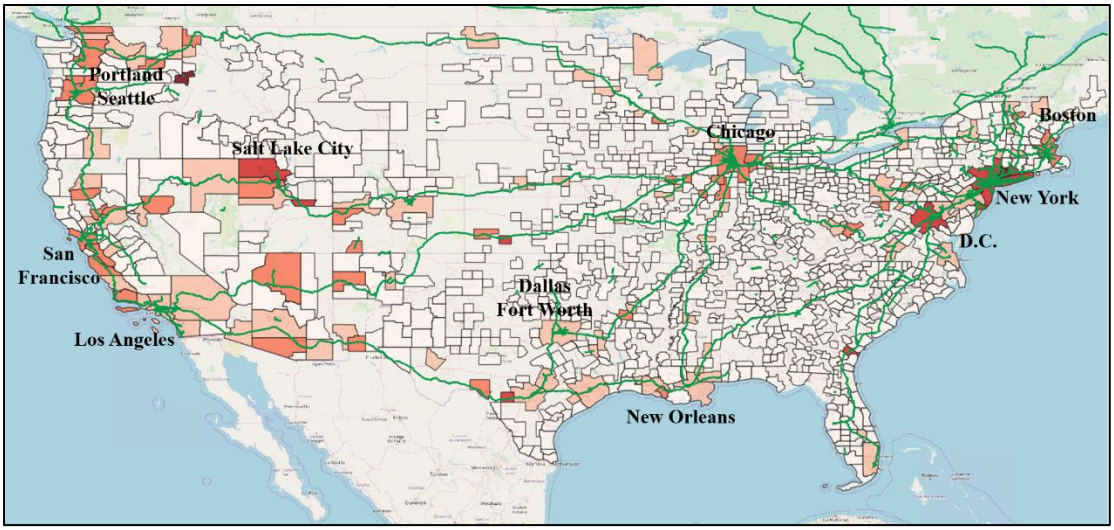


Figure 5-27. CBSA-Level Rail Mode Shares.

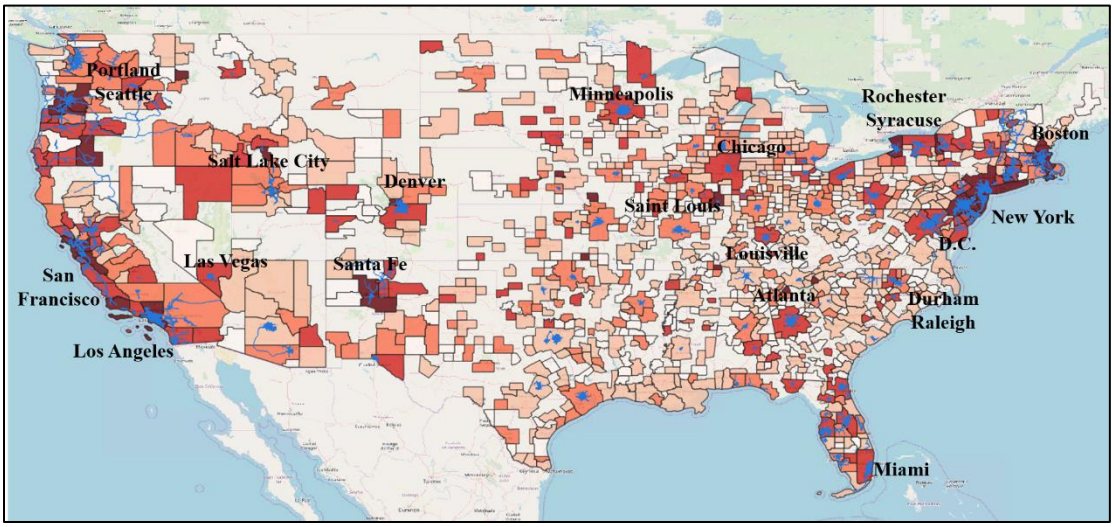


Figure 5-28. CBSA-Level Bus Mode Shares.

Chapter 6: Conclusion and Discussion

6.1 Conclusion

This study examines the state-of-the-practice applications and state-of-art-methods on processing the PCMDL data. Based on the literature review, the key research gap is identified, and a methodological framework is proposed to process the PCMDL data from raw location data to trips with imputed travel modes.

Firstly, a Spatiotemporal Density-based Spatial Clustering of Applications with Noise (ST-DBSCAN) is proposed to identify the activity locations with only PCMDL data information. Then, a novel feature construction process with multimodal transportation network information is proposed to provide inputs for the travel mode imputation models. Several machine learning methods are applied to impute travel modes examined including KNN, SVC, XGB, RF and DNN.

The framework is calibrated and compared using GPS data with user recall collected from the incenTrip application. With ground truth information, the ST-DBSCAN algorithm reaches 95% accuracy in identifying trips for different sample rates. The mode imputation results show that the Random Forest algorithm is the best model with an overall 93% accuracy to identify five travel modes, including drive, bus, rail, bike and walk.

The difference and similarity between the GPS data with user recall and LBS data are discussed to relax the models and then apply to two large-scale LBS datasets, covering the entire U.S. and the state of Maryland. For the U.S. dataset, the nationwide trip

distance and trip time distribution are compared against the 2017 NHTS. The mode shares are also compared against 2017 NHTS at a nationwide and state level. For the Maryland dataset, the statewide trip distance and trip time distribution is compared against the 2017 NHTS. And the mode share is compared against 2007/08 TPB-BMC HHTS at the county-level. The results from both the surveys and the LBS data share similar trends in terms of trip distance and trip time distributions, and mode shares.

6.2 Discussion and Future Research Directions

The limitation of this study is that the proposed methodological framework is built upon a GPS data with user recall collected in a small region. In reality, travel behavior is largely affected by geographic locations and regional transportation systems (public transit, road network etc.). Even though the comparison results for trip distance, trip time and mode shares have similar trends in comparison to the travel survey, the heterogeneity of travel behavior at different regions is not taken into account.

Apart from the training data, the features used for imputing travel modes highly rely on the multimodal transportation networks, bus trips in particular. For future research, more multimodal transportation network information can be incorporated, such as metro stations, Amtrak stations, and intercity bus stations. With more detailed information about multimodal travels, the travel mode imputation model can be potentially improved. Also, for the regions without a well-archived bus network, the bus trips can barely imputed. To decrease the dependency of transit networks, additional information such as using acceleration, stop time can be introduced.

In the two case studies, the mode shares are estimated using a small sample of LBS data, which might not be able to represent the population level travel behavior. In addition, the LBS data underrepresents the older, younger and low-income population, the results are not able to accurately capture the travel behaviors of the aforementioned population. To address these two problems, additional weighting and comparison process can be done on top of the sample results using land use, sociodemographic information.

Appendix A: Travel Mode Imputation Confusion Matrix

Table A-1. K-Nearest Neighbors Five Modes Confusion Matrix

	Rail	Bus	Bike	Walk	Drive	Recall	F1-Score
Rail	496	6	13	31	12	0.889	0.887
Bus	9	238	48	30	73	0.598	0.486
Bike	9	83	269	63	52	0.565	0.558
Walk	20	30	36	459	28	0.801	0.731
Drive	26	225	123	99	1329	0.738	0.806
Precision	0.886	0.409	0.550	0.673	0.890		

Table A-2. Support Vector Classifier Five Modes Confusion Matrix

	Rail	Bus	Bike	Walk	Drive	Recall	F1-Score
Rail	524	3	11	8	12	0.939	0.931
Bus	6	288	27	23	54	0.724	0.555
Bike	7	45	346	52	26	0.727	0.728
Walk	1	16	22	518	16	0.904	0.844
Drive	30	288	69	54	1361	0.755	0.832
Precision	0.923	0.450	0.728	0.791	0.926		

Table A-3. XGBoost Five Modes Confusion Matrix

	Rail	Bus	Bike	Walk	Drive	Recall	F1-Score
Rail	537	2	4	3	12	0.962	0.947
Bus	7	279	21	11	80	0.701	0.668
Bike	5	19	367	41	44	0.771	0.792
Walk	6	7	14	520	26	0.908	0.887
Drive	21	130	45	24	1582	0.878	0.892
Precision	0.932	0.638	0.814	0.868	0.907		

Table A-4. Random Forest Five Modes Confusion Matrix

	Rail	Bus	Bike	Walk	Drive	Recall	F1-Score
Rail	537	1	3	5	12	0.962	0.940
Bus	7	273	20	12	86	0.686	0.682
Bike	8	13	370	46	39	0.777	0.806
Walk	6	6	11	526	24	0.918	0.881
Drive	26	110	38	32	1596	0.886	0.897
Precision	0.929	0.677	0.837	0.847	0.908		

Table A-5. Deep Neural Network Five Modes Confusion Matrix

	Rail	Bus	Bike	Walk	Drive	Recall	F1-Score
Rail	471	15	21	3	48	0.844	0.880
Bus	6	328	13	10	41	0.824	0.522
Bike	4	59	353	36	24	0.742	0.749
Walk	4	25	22	508	14	0.887	0.858
Drive	27	432	57	54	1232	0.684	0.780
Precision	0.920	0.382	0.758	0.831	0.907		

Table A-6. K-Nearest Neighbors Classifier Four Modes Confusion Matrix

	Rail	Bus	NonMotor	Drive	Recall	F1-Score
Rail	502	8	35	13	0.900	0.890
Bus	9	251	59	79	0.631	0.479
NonMotor	34	148	781	86	0.745	0.742
Drive	25	243	182	1352	0.750	0.812
Precision	0.881	0.386	0.739	0.884		

Table A-7. Support Vector Classifier Four Modes Confusion Matrix

	Rail	Bus	NonMotor	Drive	Recall	F1-Score
Rail	528	3	14	13	0.946	0.932
Bus	6	304	34	54	0.764	0.562
NonMotor	11	76	919	43	0.876	0.869
Drive	30	301	98	1373	0.762	0.836
Precision	0.918	0.444	0.863	0.926		

Table A-8. XGBoost Classifier Four Mode Confusion Matrix

	Rail	Bus	NonMotor	Drive	Recall	F1-Score
Rail	535	0	9	14	0.959	0.944
Bus	7	281	21	89	0.706	0.672
NonMotor	9	29	938	73	0.894	0.902
Drive	25	128	63	1586	0.880	0.890
Precision	0.929	0.642	0.910	0.900		

Table A-9. Random Forest Four Modes Confusion Matrix

	Rail	Bus	NonMotor	Drive	Recall	F1-Score
Rail	188	0	3	2	0.974	0.964
Bus	1	88	11	25	0.704	0.667
NonMotor	2	10	328	25	0.899	0.900
Drive	6	41	22	517	0.882	0.895
Precision	0.954	0.633	0.901	0.909		

Table A-10. Deep Neural Network Four Modes Confusion Matrix

	Rail	Bus	NonMotor	Drive	Recall	F1-Score
Rail	515	11	11	21	92.29%	0.936
Bus	4	320	49	25	80.40%	0.524
NonMotor	7	52	947	43	90.28%	0.872
Drive	16	441	115	1230	68.26%	0.788
Precision	0.950	0.388	0.844	0.933		

References

1 U.S. Department of Transportation, Federal Highway Administration, 2017 National Household Travel Survey. Retrieved from: http://nhts.ornl.gov.

2 Lapham, Susan J. 1995 American Travel Survey: An Overview of the Survey Design and Methodology. 1995.

3 Zhang, L., and K. Viswanathan. The on-line travel survey manual: A dynamic document for transportation professionals. *Transportation Research Board, viewed* 17, 2013.

4 U.S. DOT Bureau of Transportation Statistics National Transit Map, 2020. https://www.bts.gov/content/national-transit-map

5 Nahmias-Biran, B. H., Han, Y., Bekhor, S., Zhao, F., Zegras, C., & Ben-Akiva, M.. Enriching Activity-Based Models using Smartphone-Based Travel Surveys. *Transportation Research Record*, 2018. *2672*(42), 280-291.

6 Wolf, J., Guensler, R., & Bachman, W.. Elimination of the travel diary: Experiment to derive trip purpose from global positioning system travel data. *Transportation Research Record: Journal of the Transportation Research Board*, 2001. 1768(1), 125-134.

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

8 Stopher, P., FitzGerald, C., & Xu, M.. Assessing the accuracy of the Sydney Household Travel Survey with GPS. Transportation, 2007. 34(6), 723-741.

9 McGowen, P., & McNally, M.. Evaluating the potential to predict activity types from GPS and GIS data. *Presented at 86th Annual Meeting of the Transportation Research Board*, Washington, D.C., 2007.

10 Battelle. Global Positioning Systems for Personal Travel Surveys: Lexington Area Travel Data Collection Test. Final Report. FHWA, U.S. Department of Transportation, 1997.

11 *2000–2001 California Statewide Household Travel Survey*. Final Report. 2002. NuStats, Austin, Tex.

12 *Kansas City Regional Travel Survey*. Final Report. NuStats, Austin, Tex, 2004.

13 Ojah, M. and Pearson, D. F.. 2006 Austin/San Antonio GPS-Enhanced Household Travel Survey. Technical Summary. 2008. Texas Department of Transportation.

14 Wolf, J., and M. Lee. Synthesis of and Statistics for Recent GPS-Enhanced Travel Surveys. Proc., *International Conference on Survey Methods in Transport*: Harmonization and Data Comparability, International Steering Committee for Travel Survey Conferences, 2008. Annecy, France.

15 *Houston-Galveston Area Council of Governments*. Draft Summary Report: 2008-09 Regional Household Activity/Travel Survey. 2009. ETC Institute.

16 *El Paso Urban Transportation Study*. Summary Report: 2010-11 Regional Household Activity/Travel Survey, 2011. ETC Institute.

17 *Wichita Falls Urban Transportation Study*. Summary Report: 2010-11 Regional Household Activity/Travel Survey, 2011. ETC Institute.

18 *Abilene Urban Transportation Study*. Summary Report: 2010-11 Regional Household Activity/Travel Survey, 2011. ETC Institute.

19 *2010–2012 Minneapolis – St. Paul Travel Behavior Inventory*. Twin Cities Metropolitan Council.

20 *2012–2013 Delaware Valley Household Travel Survey*, 2013. Delaware Valley Regional Planning Commission.

21 *Mid-Region Council of Governments 2013 Household Travel Survey*, 2014. Final Report. Westat, Rockville, Md.

22 *2014 Southern Nevada Household Travel Survey*. Final Report, 2015. Westat, Rockville, Md.

23 *Chicago Regional Household Travel Inventory*. Draft Final Report, 2007. NuStats, Austin, Tex., and GeoStats, Atlanta, Ga.

24 *2011 Atlanta, Georgia, Regional Travel Survey*. Final Report, 2011. NuStats, Austin, Tex.

25 *2010-2012 California Household Travel Survey*. Final Report Version 1.0, 2013. NuStats, Austin, Tex.

26 *Puget Sound Regional Travel Study*. Report: Spring 2014 Household Travel Survey, 2014. RSG.

27 *Puget Sound Regional Travel Study*. Report: 2015 Household Travel Survey, 2015. RSG.

28 *2017 Puget Sound Regional Travel Study.* Draft Final Report, 2017. RSG.

29 *In-The-Moment Travel Study.* Revised Report, 2015. RSG.

30 Safi, H., Assemi, B., Mesbah, M., Fereira, L., and Hickman, M.. Design and implementation of a smartphone-based system for personal travel survey: Case study from New Zealand. *Transportation Research Record: Journal of the Transportation Research Board*, 2015. vol. 2526, pp. 99–107.

31 INRIX Traffic, 2020. http://www.inrix.com/

32 Haghani, Ali, Masoud Hamedi, and Kaveh Farokhi Sadabadi. I-95 Corridor coalition vehicle probe project: Comparison of INRIX data. *I-95 Corridor Coalition* 9 2009.

33 Schrank, D., Eisele, B., & Lomax, T.. *2014 Urban mobility report: powered by Inrix Traffic Data* (No. SWUTC/15/161302-1), 2015.

34 Cui, Z., Ke, R., Pu, Z., & Wang, Y.. Deep bidirectional and unidirectional LSTM recurrent neural network for network-wide traffic speed prediction. *arXiv preprint arXiv:1801.02143*, 2018.

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

36 Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A. L.. Understanding individual human mobility patterns. *nature*, *453*(7196), 779-782, 2008.

37 Kang, C., Liu, Y., Ma, X., & Wu, L.. Towards estimating urban population distributions from mobile call data. *Journal of Urban Technology*, *19*(4), 3-21, 2012.

38 Kang, C., Ma, X., Tong, D., & Liu, Y.. Intra-urban human mobility patterns: An urban morphology perspective. *Physica A: Statistical Mechanics and its Applications*, *391*(4), 1702-1717, 2012.

39 Wang, F., & Chen, C.. On data processing required to derive mobility patterns from passively-generated mobile phone data. *Transportation Research Part C: Emerging Technologies*, 2018. *87*, 58-74.

40 Wang, F., Wang, J., Cao, J., Chen, C., & Ban, X. J.. Extracting trips from multi-sourced data for mobility pattern analysis: An app-based data example. *Transportation Research Part C: Emerging Technologies*, 2019. *105*, 183-202.

41 Wang, F., Guan, X., & Chen, C.. Assessing Impacts of Abnormal Events on Travel Patterns Leveraging Passively Collected Trajectory Data. *arXiv preprint arXiv:1911.11633*, 2019.

42 StreetLighT Data, Inc., 2020. https://www.streetlightdata.com/

43 Airsage, 2020. https://www.airsage.com/

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

45 Axhausen, K. W., Schönfelder, S., Wolf, J., Oliveira, M., & Samaga, U.. Eighty weeks of GPS-traces: approaches to enriching the trip information. *Presented at 83th Annual Meeting of the Transportation Research Board*, Washington, D.C., 2003.

46 Tsui, S. Y. A., & Shalaby, A. S.. Enhanced system for link and mode identification for personal travel surveys based on global positioning systems. *Transportation Research Record: Journal of the Transportation Research Board*, 2006. 1972(1), 38-45.

47 Bohte, W., & Maat, K..nDeriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. *Transportation Research Part C: Emerging Technologies*, 2009. 17(3), 285-297.

48 Stopher, P. R., Jiang, Q., & FitzGerald, C.. Processing GPS data from travel surveys. *2nd international colloqium on the behavioural foundations of integrated land-use and transportation models: frameworks, models and applications*, 2005. Toronto.

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

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

51 Schuessler, N., & Axhausen, K. W.. Processing raw data from global positioning systems without additional information. *Transportation Research Record: Journal of the Transportation Research Board*, 2009. 2105(1), 28-36.

52 Gong, H., Chen, C., Bialostozky, E., & Lawson, C. T.. A GPS/GIS method for travel mode detection in New York City. *Computers, Environment and Urban Systems*, 2012. 36(2), 131-139.

53 Assemi, B., Safi, H., Mesbah, M., & Ferreira, L.. Developing and validating a statistical model for travel mode identification on smartphones. *IEEE Transactions on Intelligent Transportation Systems*, 17(7), 2016. 1920-1931.

54 Patterson, Z., & Fitzsimmons, K.. Datamobile: Smartphone travel survey experiment. *Transportation Research Record: Journal of the Transportation Research Board*, 2016. *2594*(1), 35-43.

55 Gong, L., Sato, H., Yamamoto, T., Miwa, T., & Morikawa, T.. Identification of activity stop locations in GPS trajectories by density-based clustering method combined with support vector machines. *Journal of Modern Transportation*, 2015. *23*(3), 202-213.

56 Zhou, C., Jia, H., Juan, Z., Fu, X., & Xiao, G.. A data-driven method for trip ends identification using large-scale smartphone-based GPS tracking data. *IEEE Transactions on Intelligent Transportation Systems*, 2016. 18(8), 2096-2110.

57 Gong, L., Yamamoto, T., & Morikawa, T.. Identification of activity stop locations in GPS trajectories by DBSCAN-TE method combined with support vector machines. *Transportation Research Procedia*, 2018. *32*, 146-154.

58 Zhou, C., Frankowski, D., Ludford, P., Shekhar, S., & Terveen, L.. Discovering personally meaningful places: An interactive clustering approach. *ACM Transactions on Information Systems (TOIS)*, 2007. 25(3), 12.

59 Chen, W., Ji, M., & Wang, J.. T-DBSCAN: A spatiotemporal density clustering for GPS trajectory segmentation. *International Journal of Online Engineering (iJOE)*, 2014. 10(6), 19-24.

60 Ye, Y., Zheng, Y., Chen, Y., Feng, J., & Xie, X.. Mining individual life pattern based on location history. *2009 tenth international conference on mobile data management: Systems, services and middleware, 2009.* pp. 1-10..

61 Yao, Z., Zhou, J., Jin, P. J., & Yang, F.. Trip End Identification based on Spatial-Temporal Clustering Algorithm using Smartphone GPS Data (No. 19-01097), *Presented at 98th Annual Meeting of the Transportation Research Board*, Washington, D.C., 2019.

62 Stenneth, Leon, et al. "Transportation mode detection using mobile phones and GIS information." *Proceedings of the 19th ACM SIGSPATIAL international conference on advances in geographic information systems*. 2011.

63 Brunauer, R., Hufnagl, M., Rehrl, K., & Wagner, A.. Motion pattern analysis enabling accurate travel mode detection from GPS data only. In *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)* (pp. 404-411). IEEE, 2013.

64 Xiao, Guangnian, Zhicai Juan, and Chunqin Zhang. Travel mode detection based on GPS track data and Bayesian networks. *Computers, Environment and Urban Systems* 54: 14-22, 2015.

65 Broach, Joseph, Jennifer Dill, and Nathan Winslow McNeil. Travel mode imputation using GPS and accelerometer data from a multi-day travel survey. *Journal of Transport Geography* 78: 194-204, 2019.

66 Shafique, M. A., & Hato, E.. Travel mode detection with varying smartphone data collection frequencies. *Sensors*, *16*(5), 716, 2016.

67 Wang, B., Gao, L., & Juan, Z.. Travel mode detection using GPS data and socioeconomic attributes based on a random forest classifier. *IEEE Transactions on Intelligent Transportation Systems*, *19*(5), 1547-1558, 2017.

68 Highway Performance Monitoring System, 2020. https://www.fhwa.dot.gov/policyinformation/hpms.cfm

69 HERE, 2020. https://www.here.com/

70 Roth, S. B., Dematteis, J., & Dai, Y.. NHTS Weighting Report, 2017.

71 National Capital Region Transportation Planning Board, Metropolitan Washington Council of Governments. 2007/2008 TPB Household Travel Survey Technical Documentation, 2010.

72 Maryland SHRP2 C10 Implementation Assistance – MITAMS: Maryland Integrated Analysis Modelling System, 2017.

73 Birant, D., & Kut, A.. ST-DBSCAN: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering*, 2007. *60*(1), 208-221.

74 Ester, M., Kriegel, H. P., Sander, J., & Xu, X.. A density-based algorithm for discovering clusters in large spatial databases with noise. *In Kdd*, 1996.Vol. 96, No. 34, pp. 226-231.

75 Peterson, L. E.. K-nearest neighbor. *Scholarpedia*, *4*(2), 1883, 2009.

76 Cortes, C., & Vapnik, V.. Support-vector networks. *Machine learning*, *20*(3), 273-297, 1995.

77 Osuna, E., Freund, R., & Girosit, F.. Training support vector machines: an application to face detection. In *Proceedings of IEEE computer society conference on computer vision and pattern recognition* (pp. 130-136). IEEE, 1997.

78 Suykens, J. A., & Vandewalle, J.. Least squares support vector machine classifiers. *Neural processing letters*, *9*(3), 293-300, 1999.

79 Wang, L. (Ed.).. *Support vector machines: theory and applications* (Vol. 177). Springer Science & Business Media, 2005.

80 Chen, T., He, T., Benesty, M., Khotilovich, V., & Tang, Y.. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1-4, 2015.

81 Chen, Tianqi, and Carlos Guestrin. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016.

82 Michael Kearns and Leslie G. Valiant. Learning Boolean formulae or finite automata isas hard as factoring. Technical Report TR-14-88, Harvard University Aiken ComputationLaboratory, August 1988.

83 Michael Kearns and Leslie G. Valiant. Cryptographic limitations on learning Boolean formu-lae and finite automata.Journal of the Association for Computing Machinery, 41(1):67–95,January 1994

84 Liaw, A., & Wiener, M.. Classification and regression by randomForest. *R news*, *2*(3), 18-22, 2002.

85 Breiman, L.. Bagging predictors. *Machine learning*, *24*(2), 123-140, 1996.

86 Quinlan, J. R.. Induction of decision trees. *Machine learning*, *1*(1), 81-106, 1986.

87 Bengio, Y.. Learning deep architectures for AI. *Foundations and trends® in Machine Learning*, *2*(1), 1-127, 2009.

88 Schmidhuber, J.. Deep learning in neural networks: An overview. *Neural networks*, *61*, 85-117, 2015.

89 Hecht-Nielsen, R.. Theory of the backpropagation neural network. In *Neural networks for perception* (pp. 65-93). Academic Press, 1992.

90 Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P.. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357, 2002.

91 Jenks, G. F.. The data model concept in statistical mapping. *International yearbook of cartography*, *7*, 186-190, 1967.