A Data-Driven Travel Mode Share Estimation Framework based on Mobile Device Location Data

Mofeng Yang (mofeng@umd.edu)
University of Maryland, College Park https://orcid.org/0000-0002-0525-7978
Yixuan Pan
University of Maryland, College Park
Aref Darzi
University of Maryland, College Park
Sepehr Ghader
University of Maryland, College Park
Chenfeng Xiong
University of Maryland, College Park
Lei Zhang
University of Maryland, College Park

Research Article

Keywords: Travel mode share, travel surveys, machine learning, mobile device location data.

Posted Date: April 23rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-455056/v1

Version of Record: A version of this preprint was published at Transportation on August 12th, 2021. See the published version at https://doi.org/10.1007/s11116-021-10214-3.

Abstract

Mobile device location data (MDLD) contains abundant travel behavior information to support travel demand analysis. Compared to traditional travel surveys, MDLD has larger spatiotemporal coverage of the population and its mobility. However, ground truth information such as trip origins and destinations, travel modes, and trip purposes are not included by default. Such important attributes must be imputed to maximize the usefulness of the data. This paper targets at studying the capability of MDLD on estimating travel mode share at aggregated levels. A data-driven framework is proposed to extract travel behavior information from MDLD. The proposed framework first identifies trip ends with a modified 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 models. A labeled MDLD dataset with ground truth information is used to train the proposed models, resulting in a 95% recall rate in identifying trip ends and a 93% 10-fold cross-validation accuracy in imputing the five travel modes (drive, rail, bus, bike and walk) with a Random Forest (RF) classifier. The proposed framework is then applied to two large-scale MDLD datasets, covering the Baltimore-Washington metropolitan area and the United States, respectively. The estimated trip distance, trip time, trip rate distribution, and travel mode share are compared against travel surveys at different geographies. The results suggest that the proposed framework can be readily applied in different states and metropolitan regions with low cost in order to study multimodal travel demand, understand mobility trends, and support decision making.

1. Introduction

Accurate measurement of travel behavior can help agencies understand how travel demand evolves and better allocate resources in support of transportation planning processes. Traditionally, researchers and practitioners design and conduct travel surveys to obtain household- and individual-level travel behavior information, including trip origins and destinations, trip distance, trip time, trip purposes, travel modes, etc. Two of the most widely used travel surveys conducted in the United States (U.S.) are the National Household Travel Survey (NHTS, see U.S. Department of Transportation 2017) and the American Travel Survey (Lapham 1995). Methods to conduct travel surveys usually require respondents to record their daily trips with the original paper-and-pencil interview (PAPI), computer-assisted telephone interview (CATI), and computer-assisted-self-interview (CASI) (Wolf et al. 2001; Wolf 2006). However, these methods are prone to several well-known biases, such as under-reported short trips, inaccurate travel times, and travel distances (Stopher et al. 2007; McGowen and McNally 2007). Also, traditional travel surveys require complex planning and design, and large human labor and costs, and can only obtain relatively small survey samples for a limited number of cross-sections. For instance, if half of the 350 Metropolitan Planning Organizations (MPO) in the U.S. conduct travel surveys only once in a decade, it will result in $ 7.4 million per year cost (Zhang and Viswanathan 2013).

In the past two decades, along with the technological advancement in mobile sensors and mobile networks, mobile device location data (MDLD) has been growing dramatically in terms of data coverage

and data size. In the realm of transportation, the abundant individual movement information stored in MDLD has great potential to help researchers and practitioners understand the bigger picture of human travel. Compared to traditional travel surveys, MDLD has larger spatial, temporal, and population coverage and it comes from various sources, including Global Positioning Service (GPS) devices, cellular network, Bluetooth, Wi-Fi, etc. To fully take advantage of MDLD, appropriate steps and methods need to be developed to extract useful travel behavior information from MDLD (Schönfelder et al. 2002; Axhausen et al. 2003).

This study aims to develop a data-driven framework to estimate travel mode share based on MDLD. The proposed framework is trained using a labeled MDLD dataset collected from a mobile application and is further applied to two large-scale MDLD datasets. The estimated trip distance, trip time, trip rate distribution, and travel mode share are compared with travel surveys. Results suggest that the proposed framework can be readily applied in many regions with low cost to obtain travel mode share estimates and study travel trends, which can help decision-makers prioritize multimodal travel needs. The remainder of the paper is organized as follows. Section 2 reviews the literature on MDLD. Section 3 describes the proposed data-driven framework and models in detail. Section 4 introduces the data. Section 5 presents the model development results. Section 6 demonstrates the framework with two large-scale case studies. Section 7 concludes this paper and discusses the limitations and future research directions.

2. Literature Review

In this literature review, we first review the MDLDs and their applications in transportation literature. Then, we review the state-of-the-art methods for extracting trips and imputing travel modes from MDLDs.

2.1. Mobile Device Location Data and Their Applications in Transportation Literature

The earliest and most widely-used type of MDLD is the data coming from the GPS technology, where personal longitudinal location data is collected via GPS data loggers. Since the mid-1990s, researchers began investigating the possibility of using GPS data to enhance the quality of travel surveys. The initial version of the GPS data logger could only be installed in the vehicle and charged by the vehicle battery (Battelle 1997; NuStats 2002; NuStats 2004; Ojah and Pearson 2008; Wolf and Lee 2008; ETC Institute 2009; ETC Institute 2011a; ETC Institute 2011b; ETC Institute 2011c). The vehicle location was recorded seconds by seconds when the vehicle is moving (Ojah and Pearson 2006). This approach can significantly improve the spatiotemporal accuracy of travel surveys by recording the exact origin and destination as well as the trip start and end times, but it only captures vehicle trips. Later, the wearable GPS further allowed respondents to carry them such that trips traveled by other non-vehicle travel modes could also be recorded (Twin Cities Metropolitan Council 2012; Delaware Valley Regional Planning Commission 2013; Westat 2014; Westat 2015). Some travel surveys utilized both in-vehicle and wearable GPS data loggers to take advantage of both technologies (NuStats 2007; NuStats 2011; NuStats 2013). Since the GPS data can offer accurate locations of the devices, the access to individual-level trajectories is highly restricted. Therefore, the individual-level GPS data is also aggregated by private sector

companies to reveal travel demand without raising privacy concerns. For instance, INRIX Traffic collects GPS probe data from commercial vehicle fleets, connected vehicles, and mobile device applications (INRIX 2020). The data can be further aggregated into link- or corridor- level to provide a real-time estimation of traffic speed and travel time (Ali et al. 2009; Schrank et al. 2015; Cui et al. 2018).

Since mobile devices, such as smartphones and tablets, have gained in popularity, investigations into the individual-level mobility patterns have become more practical. The cellular data, which is generated through communication between cellphones and cell towers when a phone call or a text message is made by the phone (Horak 2007), has shown its great value in supporting large-scale travel demand analysis. In general, the cellular data can be categorized into Call Detail Record (CDR) and sightings (Chen et al. 2016a). Call Detail Record (CDR) data provides details on calls and messages, such as timestamp, duration, and location(s) of routing cell tower (s). Therefore, the location information of CDR data fully depends on the density of the cellular network and does not reflect the actual location of the device (Chen et al. 2016a). Similar to CDR data, sightings are also generated through communication with cell towers, but the actual location of the device is calculated via triangular calculation (Chen et al. 2016a). Both types of cellular data have been widely used in studying human mobility patterns in the past two decades. For instance, Gonzalez, et al. (2008) combined two sets of CDRs to explore individual mobility patterns; one is composed of six months of records for 100,000 randomly selected anonymous individuals and the other is a complementary dataset capturing the locations of 206 mobile phone users every two hours for one week. Further studies on human mobility have been conducted based on similar datasets (Kang et al. 2012; Kang et al. 2012; Pappalardo, et al. 2015; Song et al., 2010a; Song et al. 2010b; Çolak et al. 2016; Bachir et al. 2019; Fekih et al. 2020). Cellular data has also been widely applied to other research topics such as social network, residential location, socioeconomic level, etc. (Eagle et al. 2010; Frias-Martinez et al. 2010; Soto et al. 2011). Despite the large volume of data, cellular data is limited by its spatial and temporal resolution, which is determined by the density of cell towers and user cellphone usages (Landkmark et al. 2021). However, on the positive side, cellular data require less advanced phones and should raise less concern about the user privacy.

Another source of MDLD is the Location-based Service (LBS) data, in which spatial information is generated when a mobile application updates the device's location with the most accurate sources, based on the existing location sensors such as Wi-Fi, Bluetooth, cellular tower, and GPS (Chen et al. 2016a; Wang and Chen 2018). Compared to the CDR data, the LBS data can reflect the exact location of mobile devices and thus provide invaluable location information describing individual-level mobility patterns (Chen et al. 2016a; Gonzalez et al. 2008; Kang et al. 2012a; Kang et al. 2012b; Wang and Chen 2018; Wang et al. 2019). Lots of applications have been developed using the LBS data. For instance, a recent smartphone-enhanced travel survey conducted in the U.S. used a mobile application, rMove developed by Resource Systems Group (RSG), to collect high-frequency location data and let respondents recall their trips by showing the trajectories in rMove (RSG 2014; RSG 2015a; RSG 2015b; RSG 2017); 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 (AirSage 2020); Maryland Transportation Institute (MTI) at the University of Maryland (UMD) developed the COVID-19 Impact Analysis Platform

(*data.covid.umd.edu*) to provide insight on COVID-19's impact on mobility, health, economy and society across the U.S. (Zhang et al. 2020; Xiong et al. 2020a; Xiong et al. 2020b)

In summary, the MDLDs used in transportation literature are different in terms of spatiotemporal coverage of population and its mobility, and data quality, e.g., spatial accuracy and location recording interval (LRI) (Huang et al. 2019; Burkhard et al. 2020). The GPS data in general has the highest spatial accuracy (e.g., 10 meters) and the lowest LRI (usually 1 second), but it usually covers only a small percentage of the population, and thus cannot reflect population-level travel behavior without a statistical weighting process. Therefore, most of the GPS data are used to serve as the supplementary data sources for regional travel surveys. The cellular data and LBS data have significantly higher spatiotemporal coverage of population over the GPS data because of the large penetration rate of cellphone and mobile devices in the U.S. However, the ground truth information is usually missing and the LRI for both types of data is high and has larger variation depending on mobile device usage thus also has a larger variation (Burkhard et al. 2020). In addition, although cellular data may have higher coverage, the spatial accuracy of the data and the temporal frequency of the pins are inferior to the LBS data. This is because cellular technology relies on the density of cell towers and do not reflect the actual location of the devices. Also, cellular data is generated based on calls and messages or a network driven event which might lead to a lower number of events.

2.2. Extracting Trips from Mobile Device Location Data: State-of-the-Art Methodologies

The trip end identification algorithm for low-LRI MDLDs, i.e., GPS data, has been well-studied and used in practical applications (Huang et al. 2019). To obtain accurate trip ends, the traditional way is the rule-based trip end identification methods. This type of method designs rules and parameters based on domain knowledge. 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. The parameters used in these rules are mostly defined by domain knowledge, such as dwell time, speed, etc. (Axhausen et al. 2003; Stopher et al. 2005; Tsui et al. 2006; McGowen and McNally 2007; Du and Aultman-Hall 2007; Stopher et al. 2008; Schuessler and Axhausen 2009; Bothe and Maat 2009; Gong et al. 2012; Gong et al. 2014; Assemi et al. 2016; Patterson et al. 2016). In recent years, some researchers also leveraged the supervised machine learning models as a supplement to the rule-based methods, which classify each location point as static or moving (Gong et al. 2015; Zhou et al. 2016; Gong et al. 2018). Different clustering methods were also applied to obtain trip ends by first identifying people's activity locations from the location data (Zhou et al. 2007; Ye et al. 2009; Chen et al. 2014; Yao et al. 2019). A recent study utilized a spatiotemporal clustering method with three combined optimization models to detect trip ends (Yao et al. 2019). In recent years, there is also a special focus on deriving the trip ends from LBS data. A "Divide, Conquer and Integrate" (DCI) framework was proposed to process the LBS data to extract mobility patterns in the Puget Sound region (Wang et al. 2019). The proposed framework combined a rule-based method and incremental clustering method to handle the bi-modally distributed LBS data. The results were aggregated at census tract-level and compared with household travel surveys.

After the trip ends are identified, it is also important to impute the travel mode for each trip to obtain multimodal travel patterns. Travel mode imputation can be categorized into mainly two approaches: (1) trip-based approach; and (2) segment/point-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/point-based approach separates the trip into fixed-length segments (time or distance) or a single point and then imputes the travel mode for each segment or point (Burkhard et al. 2020). Then the segments/points with the same travel mode will be further merged to form a single-mode trip. Both previous trip-based approaches and segment/point-based approaches have used similar features in order to distinguish between different travel modes. Table 1 summarizes typical methods and features that are used for travel mode imputation.

According to the literature review done by Huang et al. 2019 and Burkhard et al. 2020, it can be observed that typical features include speed and acceleration (Stenneth et al. 2011; Gong et al. 2012; Brunauer et al. 2013; Nitsche et al. 2014; Xiao et al. 2015; Shafique and Hato 2016; Wang et al. 2017; Dabiri and Heaslip. 2018; Broach et al. 2019; Burkhard et al. 2020; Vaughan et al. 2020). Specifically, when the LRI is less than 10 seconds, the speed (speed variation) and acceleration features are more important to differentiate between different travel modes, which can be imputed solely by the data itself. When the LRI is relatively high, for instance, 30 s, additional features can be added to maintain the same level of accuracy such as real-time transit information (Stenneth et al. 2011), multimodal transportation network (Stenneth et al. 2011; Gong et al. 2012; Burkhard et al. 2020; Breyer et al. 2021), sociodemographic information (Wang et al. 2017; Vaughan et al. 2020), etc. However, most of these studies tested the algorithms using the low-LRI GPS data sample, which has frequent observations. Limited efforts have been spent on developing suitable algorithms for cellular data or LBS data that suffer from the high-LRI issue. Burkhard et al. (2020) examined the required spatial accuracy and LRI to accurately detect travel mode from the high-LRI MDLDs by subsampling the low-LRI GPS data. They concluded that the LRI should be less than a minute to ensure the travel mode imputation accuracy. Bachir et al. (2019) developed a Bayesian Inference (BI) method to separate road and rail modes from the CDR data in the Greater Paris region by leveraging the road and rail trip counts from the travel survey. Vaughan et al. (2020) trained a Deep Neural Network (DNN) model to separate drive, bus, and active modes with artificial CDR traces reconstructed from the travel survey data. The model is applied to the real-world CDR data to obtain travel mode shares. Breyer et al. (2021) developed multiple classification methods using labeled CDR data to separate only road and train mode between two OD pairs. The major limitation of these research is that either the study area is small (e.g., an OD pair or a region) or the method only separates easy-to-detect modes (e.g., Road versus Rail). As Huang et al. 2019 mentioned in their review, the supervised machine learning methods have not been fully exploited yet due to the lack of ground truth labeled data, and might be worth investigating for MDLDs, especially for the cellular data and LBS data. Besides, rather than identify easy-to-detect modes (e.g., rail versus road), it suggests including more mode categories.

This study tends to fill this gap by (1) collecting ground truth labeled LBS data to develop a supervised machine learning model that separates drive, rail, bus, walk, and bike modes; (2) developing a data-driven

travel mode share estimation framework that can be applied to large-scale LBS data and validated against the best-available ground truth data sources (Landmark et al. 2021). The proposed framework also investigates the spatial accuracy and LRI distributions of the LBS data sample and fine-tuned the algorithms considering their variations and is implemented large-scale LBS datasets. The validation efforts base the external data sources, such as household travel surveys, have proven the proposed framework efficient and versatile for applications.

Table 1
Literature Review on Travel Mode Imputation Methods.

Author	LRI	Model	Main Features	Modes	Acc.
Gong et al. 2012	/	Rules	Speed, Acceleration, Transit Stations, Transit Network	Drive, Train, Bus, Walk, Bike, Static	82.6%
Stenneth et al. 2011	30 s	RF	Speed, Acceleration, Heading change, Bus location, Transit Network	Drive, Bus, Train, Walk, Bike, Static	93.7%
Bruunauer et al. 2013	1–10 s	MLP	Speed, Acceleration, Bendiness	Drive, Bus, Train, Walk, Bike	92.0%
Xiao et al. 2015	1 s	BN	Speed, Acceleration, Trip Distance	Drive Bus, Walk, Bike, E-Bike	92.0%
Nitsche et al. 2014	1 s	DHMM	Speed, Acceleration, Direction	Drive, Bus, Motorcycle, Train, Tram, Subway, Walk, Bike	65% – 95%
Dabiri and Heaslip. 2018.	1–5 s	CNN	Speed, Acceleration, Jerk, Bearing Rate	Drive, Bus, Train, Walk, Bike	84.8%
Bachir et al. 2019	/	BI	Road and Rail Trip Counts	Road, Rail	/
Vaughan et al. 2020	/	DNN	Speed, Trip Distance, Land Use, Time of Day	Drive, Bus, Active (Walk, Bike)	87%
Burkhard et al. 2020	1 s subsampled to 5 min	KNN, RF etc.	Speed, Public Transport Stops and Lines	Drive, Train, Tram, Bus, Walk, Bike	/
Breyer et al. 2021	/	KNN etc.	Road and Train Route Geometry	Road, Train	95.5%

** RF: Random Forest; MLP: Multi-Layer Perceptron; BN: Bayesian Network; DHMM: Discrete Hidden Markov Model; CNN: Convolutional neural Network; BI: Bayesian Inference; DNN: Deep Neural Network.*

3. The Data-driven Travel Mode Share Estimation Framework

Figure 1 shows the proposed data-driven framework. On the left is the Model Development pillar, wherein a dedicated ground-truth data collection of labeled and mode-specific trips and trajectories is conducted in order to train the trip end identification algorithm and travel mode imputation model. These trained models are then applied to the Model Application pillar on the right. The Model Application generates trip rosters with imputed travel modes for the unlabeled MDLD datasets in the application contexts. Finally, a validation process compares the aggregated mode share, as well as other statistics, with travel surveys before the data products are deemed useful and applicable for any transportation planning applications.

3.1. Trip End Identification

Considering a person's daily travel, it is very common that he or she makes multiple stops for different trip purposes. As illustrated in Fig. 2, these stops are categorized into two categories, namely Activity Stops (AS) and Non-Activity Stops (NAS). ASs represent stops where actual activities take place, such as home, workplace, restaurant, shopping mall, etc. NASs represent stops 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 are considered as actual trip ends and the trajectory between two consecutive ASs is considered as a valid trip.

The first step is to identify all stop points including all ASs and NASs. The Spatiotemporal Density-based Spatial Clustering of Applications with Noise (ST-DBSCAN) (Birant and Kut 2007) is applied to fulfill this step. The ST-DBSCAN is an extended version of the traditional DBSCAN algorithm (Ester et al. 1996) with consideration of both spatial and temporal constraints. The temporal constraint was able to handle the scenarios when a person visits the same location multiple times per day, i.e., home, work. Three thresholds are defined for the ST-DBSCAN used in this study: (1) the spatial threshold s: it represents the distance falling within the activity distance range; (2) the temporal threshold t: it represents the minimum duration of an activity; and (3): the minimum neighbor's m: it represents the minimum number of location points to form a cluster. Details of ST-DBSCAN can be found in Birant and Kut (2007) and Ester et al. (1996).

With all stop points identified, the second step is to validate and distinguish between ASs and NASs. Two parameters are proposed: (1) s_{act}: maximum activity distance range. If the distance between two consecutive clusters stayed within s_{act}, it implies that these two clusters might still belong to the same activity, and the location points falling within these two clusters would be labeled as activities, otherwise, a trip will be generated. (2) t_{act}: minimum activity duration. t_{act} is used to compare with the time lag calculated between the last observed location point and the first observed location point of a cluster. If the time lag is shorter than t_{act}, it implies no activity happens. The cluster will be flagged and all the locations within this cluster will be considered as trip waypoints of the previous trip. This can occur when waiting at traffic lights, encountering traffic congestion, etc. Otherwise, the cluster will be identified as an activity.

3.2. Travel Mode Imputation with Machine Learning Models

This study proposes machine learning models to impute five travel modes (drive, bus, rail, bike and walk) and four travel modes (drive, bus, rail, and non-motorized) from trips identified from the previous step. Five machine learning models 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). The five machine learning modes are implemented using the scikit-learn package in Python (Pedregosa et al. 2011). A detailed introduction of these models can be found in Appendix I.

Feature set construction directly affects the model performances. Three types of features, LRI feature, trip features, and multimodal transportation network features, are constructed from MDLD, as shown in Table 2. The LRI feature, represented by the average number of records per minute, indicates the location service usage during a trip. The trip features can show the characteristics of each trip, including trip distance, origin-destination distance, trip time, average speed, minimum speed, maximum speed, median speed, and 5, 25, 75, 95- percentile speed. The multimodal transportation network features are important to distinguish between different travel modes (Bohte and Maat 2009; Gong et al. 2018). Here, a 50-meter buffer for the multimodal transportation networks (rail, bus, and drive), and bus stops are generated to obtain the percentage of location points for each trip that fall within each buffer, respectively. It should be noted that since the accuracy of these multimodal network features is largely affected by the spatial accuracy of the location data, we used only location data with spatial accuracy less than 50 meters to calculate these features to obtain more accurate estimates. Accelerator meter data is also useful in travel mode imputation. However, it was not taken into account because such information is generally not available in large-scale MDLD datasets purchased from third-party vendors.

Table 2
Features Constructed from Mobile Device Location Data

Features	Unit
Location Recording Interval Feature	
Average # of records per minutes	frequency
Trip Features	
Trip distance	meters
Origin-Destination distance	meters
Trip time	minutes
Max., Min., Avg, Med., 5-, 25-, 75-, 95- percentile speed	meter / second
Multimodal Transportation Network Features (location data with accuracy < 50 meters)	
% of location points fell within 50 meters of rail network	percentage
% of location points fell within 50 meters of bus network	percentage
% of location points fell within 50 meters of drive network	percentage
% of location points fell within 50 meters of bus stops	percentage

4. Data

4.1. Mobile Device Location Data

This study uses three MDLD datasets that can be categorized into LBS data. The first LBS dataset is collected from a mobile application called incenTrip (*incentrip.org*). incenTrip was developed by Maryland Transportation Institute (MTI) at the University of Maryland (UMD) to nudge travel behavior changes by providing real-time dynamic incentives in the Washington Metropolitan Area (Xiong et al. 2019). incenTrip collects location data using the Google Maps API (Android) or Apple Core Location (iOS) with a set of pre-defined fixed LRIs and inserts the data into PostgreSQL in the Amazon Web Services (AWS) to ensure privacy protection. The LRI of the incenTrip data is set to be 1–15 seconds with three considerations: 1) help preserve battery of the mobile devices, and at the same time fully capture users daily travel behavior; 2) ensure the capability of data for travel mode imputation, as the literature suggests, reducing the LRI of the data will decrease the travel mode imputation accuracy (Shafique and Hato 2016); and 3) improve similarity with large-scale LBS datasets. Each record of the raw location data includes a unique device identifier, latitude, longitude, and timestamp information. The proposed framework would then be applied to identify trips, impute the corresponding travel modes, and then update the trip records with imputed modes in the database. This study uses the incenTrip app from March 2019 to January 2020 for a

dedicated ground-truth MDLD data collection. During the 10-month period, fifteen designated respondents of all the incenTrip users were hired to travel with incenTrip and record detailed information for each trip daily, including the start date, start time, end date, end time, origin street address, destination street address, travel time and travel mode. The trip identification algorithm is calibrated by comparing their travel diaries and the algorithm outputs. In addition, we also collected the trips with confirmed travel modes by other users. As a result of this data collection effort, a total number of 12,688 ground-truth trip records with travel mode labels were obtained from 410 users for the subsequent travel mode imputation model training process. Figure 3 visualizes the trajectories of these trips in different colors by travel modes.

Another two LBS datasets are obtained from one of the leading data vendors in the U.S. Similar to the description provided in the literature, the data is passively collected through mobile device applications. Each observation includes time of the event in Unix epoch time format, an anonymized hashed identification number (ID), latitude and longitude coordinates in decimal version, location accuracy of each observation in meter, and the time zone offset associated with the position of each observation. In this study, the two LBS datasets are extracted with different spatial, temporal, and population coverages. Table 3 describes these two datasets. Dataset I covers the Baltimore-Washington metropolitan area, including the State of Maryland, District of Columbia (D.C.), and Northern Virginia. It has a temporal coverage of one typical weekday, Sept. 12nd in 2017. A total of 474,634 unique devices observed in this area are considered. Dataset II expands the spatial coverage to the U.S. It has a temporal coverage of seven days from Aug. 1st to Aug. 7th in 2017. Since the focus of this study is to demonstrate the transferability and capability of our proposed framework on deriving travel mode shares, 3% of the total number of devices observed is randomly sampled and used for this study in order to reduce the computation burden, including 266,149 unique devices. Figure 4 (a) and (b) visualizes the raw location points for Dataset I and Dataset II respectively, where the lighter area means more concentrated data.

Table 3
Location-Based Service Data Description

Dataset	Spatial Coverage	Temporal Coverage	Sample Rate	Sampled Numbers
I	Baltimore-Washington metropolitan area	Sept. 12nd, 2017	100%	474,634
II	the United States	Aug. 1st – Aug. 7th, 2017	3%	266,149

4.2. Multimodal Transportation Networks

This study also collects the multimodal transportation network data including drive, bus, rail networks, and bus stop locations to construct network-related features. The drive network is collected from the

Highway Performance Monitoring System (HPMS) (FHWA 2020) that includes national freeway and arterial roads in the U.S. The national bus and rail network and the bus stops data are collected from the United States Department of Transportation (U.S. DOT) Bureau of Transportation Statistics (BTS) National Transit Map (NTM) (U.S. DOT BTS 2020). Figure 5 illustrates the multimodal transportation networks used in this study.

4.3. Travel Surveys

Two travel surveys are used in this study for comparison purposes: NHTS 2017 and 2007/08 TPB-BMC Household Travel Survey (HHTS). NHTS 2017 is a national-level travel survey conducted by USDOT Federal Highway Administration (FHWA), collecting travel behavior data from U.S. residents. The NHTS 2017 includes a total number of 129,696 households covering all 50 states and the District of Columbia, including trip origin and destinations, trip time, trip purposes, and travel modes (U.S. DOT 2017). The 2007/2008 TPB-BMC HHTS is conducted by Transportation Planning Board (TPB) and Baltimore Metropolitan Council (BMC) in Baltimore and Washington regions from February 2007 to March 2008 using the same survey designs (MWCOG 2010). This survey covered nearly 14,000 households and can provide travel mode shares at Traffic Analysis Zone (TAZ) level. In our case study, we aggregated the travel mode shares at the county level using the trip origins in the travel survey and further compare with the LBS estimates.

5. Model Development

5.1. Trip End Identification Result

Five parameters of the proposed ST-DBSCAN are calibrated using the incenTrip data: 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}. Since s determines the distance range of a stop, increasing the value of s would identify less stop points since more location points would be clustered. To ensure all the stop points are captured including traffic congestions and waiting at a traffic light for both vehicle and pedestrian, four constraints are added as shown below:

$$t \geq n \cdot f$$
$$t_{act} \geq n \cdot f$$
$$s_{act} \geq s$$
$$n \cdot f \geq s/v$$

where v is the average walking speed, here we consider 1 m/s; f is location recording interval. Consider the real-world scenario when a person stops, it is intuitive to set the s value to be relatively small. 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 LRI the corresponding range for other parameters could be calculated. Table 4 shows the calibrated parameters used in the case study for each LRI.

Figure 6 shows the trip end identification result. The “Reported Trips” represents the trips that are reported in the respondents’ travel diaries; the “Matched Trips” represents how many of the “Reported Trips” are identified from the data; the “All Identified Trips” represents all trips identified from the data, including the matched trips and the underreported trips that are not recorded in respondents’ travel diary; and the “Hit-Ratio” (or recall rate) is the value of “Matched Trips” divided by “Reported Trips”. It should be noted that for the 1-s LRI data, only 23 reported trips are collected due to the short testing period. Over 90% of reported trips can be identified for each LRI, with the overall Hit-Ratio equals to 94.5%. In addition, about 15–35% of the underreported trips are identified and confirmed by testers. Capturing these underreported trips help produce more detailed travel patterns for each respondent.

Table 4
Calibrated Parameters for Each Location Recording Interval.

LRI (s)	s (m)	t (s)	n	s_{act} (m)	t_{act} (s)
1	50	100	50	100	300
2	50	200	25	100	300
5	50	500	15	100	300
15	50	600	10	100	300

5.2. Travel Mode Imputation Result

Five machine learning models, KNN, SVC, XGB, RF, and DNN are used to impute travel modes. A total of 6,064 drive trips, 1,824 rail trips, 1,403 bus trips, 1,496 bike trips, and 1,901 walk trips are collected from the incenTrip. 70% of the data is used for training and 30% of the data is used for testing. The Synthetic Minority Over-Sampling Technique (SMOTE) is then applied to the training data to address the imbalanced sample problem, where the minority class from the existing samples is synthesized (Chawla et al. 2002). For each machine learning method, the randomized search approach is used to fine-tune the model. Detailed hyperparameters can be found in Appendix I. During the model training process, 10-fold cross-validation (CV) is conducted to fine-tune the model parameters using 70% training data. The fine-tuned models are then applied to the 30% testing data and the F1 scores are calculated using the equations as shown below:

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

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

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

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

Table 5 compares the training accuracy with 10-fold CV and the F1 scores calculated after applying the models to the testing data. It can be seen that RF achieved the highest CV-accuracy in both four and five modes models. The bus mode has the least prediction accuracy than the other modes. One possible reason might be the similarity between the drive trips and bus trips. As shown in Fig. 3 (a) and (b), the drive trips cover most interstate highways, major arterials while the bus trips will only follow the pre-designed bus routes and will stop at the bus stops. However, when it comes to the roadway segments where bus routes exist, the moving patterns of passenger cars and buses highly depend on the traffic conditions and thus could share similar spatiotemporal characteristics, which requires high-quality data and complex method to distinguish with each other (Nguyen and Armoogum 2020). Therefore, in such roadways, especially in urban settings, the general level of LBS data quality might not be sufficient to capture the differences in their moving patterns, e.g., distinguishing between the short and frequent stops due to traffic congestion and bus stops. The RF model's feature importance can also be found in Appendix I.

Table 5
Model Performance Comparison.

		KNN	SVC	XGB	RF	DNN
Four Modes	Drive	81.4%	82.1%	88.9%	88.7%	85.4%
	Rail	87.2%	90.6%	91.3%	90.8%	89.2%
	Bus	50.1%	52.2%	64.3%	65.2%	57.3%
	NonMotor	76.4%	85.4%	88.8%	88.8%	86.5%
10-Fold CV Accuracy		86.0%	86.3%	93.0%	93.3%	86.5%
Five Modes	Drive	79.6%	81.9%	88.8%	88.7%	83.7%
	Rail	86.7%	90.5%	91.3%	91.2%	90.2%
	Bus	47.7%	53.1%	64.4%	64.8%	56.3%
	Bike	50.0%	68.8%	77.3%	78.6%	72.5%
	Walk	69.0%	82.4%	87.4%	87.5%	85.4%
10-Fold CV Accuracy		85.5%	85.4%	93.0%	93.6%	85.0%

** The 10-Fold CV Accuracy are obtained from 70% training data and F1 scores for each mode are calculated using the 30% testing data.*

6. Case Studies

6.1. Comparison between the incenTrip Data and the Two LBS Datasets

Before we apply the calibrated framework to the two large-scale LBS datasets, the spatial accuracy and the LRI of these three datasets need to be compared. Figure 7 (a) shows the spatial accuracy of the three datasets after removing the data with an accuracy greater than 100 meters (see Appendix II.A. for more details). The incenTrip dataset has the best spatial accuracy among the three datasets, with more than 80% of the data's spatial accuracy less than 50 meters, whereas the two LBS datasets show a bimodal distribution, with the second peak locates around 70 meters. In general, the spatial accuracy of these three datasets is of high quality because, for all three datasets, around 90% of the data has a spatial accuracy of less than 100 meters. Figure 7 (b) only shows the LRI distribution for the two large-scale LBS datasets, since the incenTrip dataset is collected with pre-defined LRI, where a bimodal distribution can be observed (Wang et al. 2019). For each dataset, more than 75% of the data has LRI less than 15-s. Therefore, with consideration of both spatial accuracy and LRI, we applied the framework calibrated by the incenTrip data with 15-s LRI to estimate trips and travel mode shares for the two LBS datasets. In addition, since we observed that the second peak of the bimodal distribution locates around LRI = 120 s, two further relaxations are made in order to relax the restrictions of activity cluster identification: (1) the

temporal threshold t is relaxed from 600 s to 1800 s; (2) the minimum neighbors n is relaxed from 10 to 5. Increasing the temporal threshold t and decreasing the minimum neighbors n can help capture a comparable number of activity clusters for LBS data with small LRI. Therefore, as shown in the previous calibration results (Table 4), data with high LRI requires larger t and smaller n. the final parameters used for the two case studies are s = 50 m, t = 1800 s, n = 5, s_{act} = 100 m and t_{act} = 300 s. More analysis regarding the spatial accuracy and LRI of the LBS datasets used in this study can be found in Appendix II.A. and Appendix II.B.

6.2. Case Study I: Baltimore-Washington Metropolitan Area Dataset

In the first case study, the proposed framework is applied to the LBS data observed in the Baltimore-Washington metropolitan area, covering the state of Maryland, D.C., and Northern Virginia. The trip distance, trip time and trip rate distribution are firstly compared with the 2007/08 TPB-BMC HHTS. The travel mode share is then compared at statewide- and county-level. A visualization of bus and rail travel distribution is provided at the census tract-level.

Figure 10 visualizes the census tract-level rail and bus travel mode shares using Jenks natural breaks optimization (Jenks 1967), with the deeper color representing higher mode share. For both D.C. and Baltimore city, the travel mode share distribution follows the geographical layouts of rail and bus networks. Also, since D.C. has denser rail and bus networks, the relative mode share is higher than that in Baltimore City.

6.3. Case Study II: the U.S. National Dataset

In the second case study, the proposed framework is applied to the LBS data observed in the entire U.S. for a week, with 3% of the observed devices randomly sampled. The trip distance, trip time, trip rate distribution, and travel mode share are compared against the NHTS 2017 at the nationwide- and state-level. A visualization is provided at CBSA-level.

The LBS estimates of trip distances, trip times, and trip rates distributions are compared with the latest NHTS 2017, as shown in Fig. 11 (a), (b), and (c). Compared to the results in Case Study I, the difference observed in the long-distance part is smaller. The main reason is that NHTS has larger spatial coverage of households and captures more longer-distance and longer-duration trips than 2007/08 TPB-BMC HHTS. Another possible reason for the difference observed in the overall trip distance distribution is that NHTS 2017 calculated trip distance using the network shortest path from Google API, which is not always representative of the real path taken. The discussions about trip time and trip rate distributions in Case Study I can also account for the differences observed here in the national comparison. Though discrepancies still exist, an overall satisfactory match is achieved. Different from Case Study I, since a week of data is used here, the day of week variation is also presented. It should be noted that the data is stored by (Coordinated Universal Time) UTC date so only six full local days data can be extracted. As

shown in Fig. 12, the differences between weekends and weekdays are captured, where the peak hours cannot be seen on Saturday and Sunday.

Before applying the travel mode imputation model, the air trips are firstly filtered out with a heuristic rule using three trip features: 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 is larger than 100 mph, the trip time is larger than 1 hour and the trip distance is larger than 100 miles, then this trip is labeled as an air trip. Figure 13 shows the heat map of air trips' origins, where the depth of the color represents the number of air trips originated from the closest airport. It can be observed that almost all the major airports are captured.

Figure 15 visualizes the CBSA-Level rail and bus travel mode share. The CBSAs with high bus and rail travel mode share estimates are those who have well-developed bus and rail networks. For instance, D.C., New York, Boston, and San Francisco have higher rail mode share than the other CBSAs. For bus mode share, a similar trend is observed too, where CBSAs with denser bus networks have higher bus mode share.

7. Conclusions And Discussions

This study reviews the state-of-the-practice applications and state-of-the-art methods for extracting travel behavior information from MDLDs. Based on the literature review, the key research gap is identified, and a data-driven framework is proposed to estimate travel mode shares from MDLDs. The proposed framework reaches a 95% recall rate in identifying trip ends and a 93% 10-fold cross-validation accuracy in identifying five travel modes with the RF model. The developed framework is applied to two large-scale LBS datasets, covering the Baltimore-Washington metropolitan area and the U.S. with different spatial, temporal, and population coverage. The trip distance, trip time, and trip rate distribution, and travel mode share estimated from the two LBS datasets are compared with travel surveys for a comprehensive validation. The comparison results suggest that the proposed framework performs robustly in both geographical regions, indicating a good transferability.

One major challenge in the travel mode imputation is to distinguish different travel modes in urban settings, especially between drive and bus. This is also the reason why we solicited our training samples in the Washington Metropolitan Area. In the meantime, the multimodal networks are simpler in most rural areas: the rail network is sparse, the bus service is limited, etc. From our comparison with travel surveys at the county/state level, it shows that the current model trained with our incenTrip dataset could achieve satisfactory estimation in other contexts or regions.

Another focus of this study is to comprehensively compare the LBS estimates of trip distance, trip time, trip rate and travel mode share with a regional travel survey, 2007/08 TPB-BMC HHTS, and a national travel survey, NHTS 2017. The major differences shown by the comparisons between the LBS estimates and survey results imply that the decision-makers should be very careful when handling the LBS data to support travel surveys and other transportation applications. Firstly, it is necessary to evaluate different trip distance calculation methods based on LBS data and select the one that fits the research or application needs the best. Secondly, due to the technology limitation, the LBS data, or in general, MDLDs, only captures part of the daily trips of a device while the survey usually captures all trips in a typical travel day. This variability in the observed duration among devices might also result in capturing more long distance/duration trips from active travelers, such as long-distance travel for leisure or business purposes or long-distance commute. Additional trip-level adjustment by distance bands might be useful to reduce such bias in the LBS data. Lastly, because of the signal loss and urban canyon effect, especially in major cities and metropolitan areas, the bus and rail travels might be hard to captured or imputed from the LBS data, resulting in an overall underestimation of travel mode share as compared to travel survey. Future research is also suggested to adopt trip-level adjustment by different travel modes based on control total data such as the National Transit Database (NTD) (Chu 2010).

The limitations of this study are summarized and discussed into three aspects: training data, validation data, and sample bias.

- **Training Data**: The proposed framework is calibrated and trained based on samples collected in the Washington Metropolitan Area with the incenTrip application. Even though the general travel mode share statistics are consistent with the travel surveys, in the real world, the travel behavior might be different from region to region, resulting in biased travel mode share estimation. In future research, enriching the training dataset that could cover different regions, such as the GPS-enhanced travel survey dataset available in Transportation Secure Data Center (TSDC) at National Renewable Energy Lab (NREL), might help improve the performance of our proposed framework by considering the travel behavior heterogeneity. In addition, the proposed framework relies on multimodal transportation networks, including drive, rail and bus. For regions without well-maintained transportation networks, it could be hard to capture the rail/bus travel. To decrease the dependency on multimodal transportation networks, additional information such as acceleration and stop time can be potentially considered.
- **Validation Data**: The proposed framework provides a general way to estimate travel mode shares at different geographies, with drive mode well captured, bus and rail mode slightly underestimated and non-motorized mode slightly overestimated. To further improve the performance of the proposed framework and capture finer-level multimodal travel trends, additional data (i.e., transit ridership data, station-based metro passenger volume data) could be collected as external validation sources and control totals. In addition, the current heuristic rule filters air trips according to domain knowledge, such as thresholds of average travel speed, trip time, and trip distance, which could be refined with a

ground-truth data collection for long-distance travels. A similar data collection effort to the procedure done by this paper could address the limitation and is on our research agenda.

- **Sample Bias**: In the two case studies, the travel mode share is estimated from the LBS data with a sample of the population, which might not be able to represent the population-level travel behavior. Also, the LBS data could not capture the travel behaviors of the population without mobile devices, which might yield an underrepresentation of the younger, elder, and low-income population. To address these two problems, an additional weighting and validation process can be done on top of the sample results using land use and sociodemographic information. In addition, 3% of the devices are randomly sampled from the Dataset II to reduce the computation load. Future studies are suggested to leverage cloud computing techniques in order to take advantage of the entire dataset to increase the sample size.

In summary. our proposed framework is able to provide a timely and continuous measurement of travel trends and travel mode shares as a supplement to travel surveys. To achieve better estimates in practice, our proposed framework should be firstly calibrated using the travel surveys and then be applied to the MDLD datasets for preliminary investigation of the latest travel trends and mode shares. Although the proposed framework is not ready to completely replace travel surveys, it could help the government and local agencies refine the design of their travel surveys after prioritizing their data needs and before investing time and money in actual implementation. In addition, the proposed framework can also be applied to other realms, such as business development and public health. For instance, during the COVID-19 pandemic, a handful of research utilized the MDLDs to derive the travel statistics, i.e., trip rate and inflow, and study their correlation with the new COVID-19 infections (Xiong et al. 2020a; Xiong et al. 2020b; Pan et al. 2020). By applying our proposed framework to the MDLDs, the correlation between multimodal travel and the spread of COVID-19 can be studied to support governments decision makings on transit or airline operations.

Declarations

Funding

The research is partially financially supported by a USDOT Federal Highway Administration (FHWA) Exploratory Advanced Research (EAR) project entitled “Data Analytics and Modeling Methods for Tracking and Predicting Origin-Destination Travel Trends based on Mobile Device Data” (Award No. 693JJ31750013). The opinions in this paper do not necessarily reflect the official views of USDOT or FHWA.

Conflict of interest

The authors declare that they have no conflict of interest.

Availability of data and material

Not applicable.

Code availability

Not applicable.

Authors' contributions:

M.Y., Y.P., and L.Z. designed research; M.Y., A.D. and S.G. processed data; M.Y., Y.P., and C.X. developed the algorithms and models; M.Y., A.D. and S.G. conducted case studies; M.Y., Y.P., A.D., S.G., C.X., L.Z. wrote the paper.

References

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

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

2010–2012 Minneapolis – St. Paul Travel Behavior Inventory. Twin Cities Metropolitan Council, (2012).

2011 Atlanta, Georgia, Regional Travel Survey. Final Report. NuStats, Austin, Tex, (2011).

2012–2013 Delaware Valley Household Travel Survey. Delaware Valley Regional Planning Commission, (2013).

2014 Southern Nevada Household Travel Survey. Final Report. Westat, Rockville, Md, (2015).

2017 Puget Sound Regional Travel Study. Draft Final Report. RSG, (2017).

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

Airsage. https://www.airsage.com/, (2020).

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

Bachir, D., Khodabandelou, G., Gauthier, V., El Yacoubi, M. and Puchinger, J. Inferring dynamic origin-destination flows by transport mode using mobile phone data. *Transportation Research Part C: Emerging Technologies*, 101, pp.254-275. (2019).

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

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

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

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*. 17(3), 285-297, (2009).

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

Breyer, N., Gundlegård, D. and Rydergren, C.. Travel mode classification of intercity trips using cellular network data. *Transportation Research Procedia*, 52, pp.211-218. (2021).

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).

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).

Burkhard, O., Becker, H., Weibel, R. and Axhausen, K.W.. On the requirements on spatial accuracy and sampling rate for transport mode detection in view of a shift to passive signalling data. *Transportation Research Part C: Emerging Technologies*, 114, pp.99-117. (2020).

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).

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

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

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*. 68, 285-299, (2016a).

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. (2016b).

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

Chu, X.. A guidebook for using automatic passenger counter data for national transit database (NTD) reporting (No. NCTR778-03, FDOT BDK85 977-04). National Center for Transit Research (US) (2010).

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

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).

Dabiri, S. and Heaslip, K.. Inferring transportation modes from GPS trajectories using a convolutional neural network. *Transportation research part C: emerging technologies*, *86*, pp.360-371. (2018).

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*, 41(3), 220-232, (2007).

Eagle, N., M. Macy and R. Claxton. Network Diversity and Economic Development. *Science* Vol. 328, No. 5981, pp. 1029-1031. (2010).

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

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

Fekih, M., Bellemans, T., Smoreda, Z., Bonnel, P., Furno, A. and Galland, S.. A data-driven approach for origin–destination matrix construction from cellular network signalling data: a case study of Lyon region (France). *Transportation*, pp.1-32. (2020).

Frias-Martinez, V., J. Virseda, A. Rubio and E. Frias-Martinez. Towards Large Scale Technology Impact Analyses: Automatic Residential Localization from Mobile Phone-Call Data. *Proceedings of the 4th ACM/IEEE international conference on information and communication technologies and development*, ACM. (2010).

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, (2012).

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*. 138, 557-565, (2014).

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*. 23(3), 202-213, (2015).

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*. 32, 146-154, (2018).

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

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

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

HERE. https://www.here.com/, (2020)

Highway Performance Monitoring System, Federal Higway Administration. https://www.fhwa.dot.gov/policyinformation/hpms.cfm, (2020).

Horak, Ray. *Telecommunications and data communications handbook*. John Wiley & Sons, (2007).

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

Hu, Patricia S., and Timothy R. Reuscher. Summary of travel trends: 2001 national household travel survey. (2004).

Huang, H., Cheng, Y. and Weibel, R.. Transport mode detection based on mobile phone network data: A systematic review. *Transportation Research Part C: Emerging Technologies*, *101*, pp.297-312. (2019).

INRIX Traffic. http://www.inrix.com/, (2020).

In-The-Moment Travel Study. Revised Report. RSG, (2015a).

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

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, (2012a).

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, (2012b).

Kansas City Regional Travel Survey. Final Report. NuStats, Austin, Tex, (2004).

Landmark, A.D., Arnesen, P., Södersten, C.J. and Hjelkrem, O.A.. Mobile phone data in transportation research: methods for benchmarking against other data sources. *Transportation*, pp.1-23. (2021).

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

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

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).

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

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, (1994)

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

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

Nguyen, M. H., and Armoogum, J.. Hierarchical process of travel mode imputation from GPS data in a motorcycle-dependent area. *Travel behaviour and society*, 21, 109-120 (2020).\

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

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

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).

Pan, Y., Darzi, A., Kabiri, A., Zhao, G., Luo, W., Xiong, C., and Zhang, L.. Quantifying human mobility behaviour changes during the COVID-19 outbreak in the United States. *Scientific Reports*, 10(1), 1-9 (2020).

Pappalardo, L., F. Simini, S. Rinzivillo, D. Pedreschi, F. Giannotti and A.-L. Barabási. Returners and Explorers Dichotomy in Human Mobility. *Nature communications*. Vol. 6,pp. 8166. (2015).

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J.. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12, 2825-2830, (2011).

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

Puget Sound Regional Travel Study. Report: Spring 2014 Household Travel Survey. RSG, (2014).

Puget Sound Regional Travel Study. Report: 2015 Household Travel Survey. RSG, (2015b).

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

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*. vol. 2526, pp. 99–107, (2015).

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

Schönfelder, Stefan, et al. Exploring the potentials of automatically collected GPS data for travel behaviour analysis: A Swedish data source. *Arbeitsberichte Verkehrs-und Raumplanung* 124 (2002).

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

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

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

Song, C., T. Koren, P. Wang and A.-L. Barabási. Modelling the Scaling Properties of Human Mobility. *Nature Physics*. Vol. 6, No. 10, pp. 818. (2010a).

Song, C., Z. Qu, N. Blumm and A.-L. Barabási. Limits of Predictability in Human Mobility. *Science*. Vol. 327, No. 5968, pp. 1018-102. (2010b).

Soto, V., V. Frias-Martinez, J. Virseda and E. Frias-Martinez. Prediction of Socioeconomic Levels Using Cell Phone Records. *International Conference on User Modeling, Adaptation, and Personalization*, Springer. (2010).

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).

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*. Toronto, (2005).

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

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

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

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*. 1972(1), 38-45, (2006).

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

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

U.S. Department of Transportation, Bureau of Transportation Statistics, *Transportation Statistics Annual Report 2020*. Washington, DC. https://doi.org/10.21949/1520449. (2020).

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

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).

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

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*. 105, 183-202, (2019).

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

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*. 1768(1), 125-134 (2001).

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

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. Annecy, France (2008).

Xiao, G., Juan, Z., and Zhang, C.. Travel mode detection based on GPS track data and Bayesian networks. *Computers, Environment and Urban Systems* 54: 14-22, (2015).

Xiong, C., Shahabi, M., Zhao, J., Yin, Y., Zhou, X., and Zhang, L.. An integrated and personalized traveler information and incentive scheme for energy efficient mobility systems. *Transportation Research Part C: Emerging Technologies* (2019).

Xiong, C., Hu, S., Yang, M., Luo, W., and Zhang, L.. Mobile device data reveal the dynamics in a positive relationship between human mobility and COVID-19 infections. *Proceedings of the National Academy of Sciences*, 117(44), 27087-27089 (2020a).

Xiong, C., Hu, S., Yang, M., Younes, H., Luo, W., Ghader, S. and Zhang, L.. Mobile device location data reveal human mobility response to state-level stay-at-home orders during the COVID-19 pandemic in the USA. *Journal of the Royal Society Interface*, 17(173), p.20200344. (2020b).

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).

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*. pp. 1-10, (2009).

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

Zhang, L., Sepehr G., Michael L. P., Chenfeng X., Aref D., Mofeng Y., Qianqian S., AliAkbar K., and Songhua H.. An interactive COVID-19 mobility impact and social distancing analysis platform. *medRxiv* (2020).

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

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*. 18(8), 2096-2110, (2016).

Çolak, S., A. Lima and M. C. González. Understanding Congested Travel in Urban Areas. *Nature communications*. Vol. 7, pp. 10793. (2016).

Figures

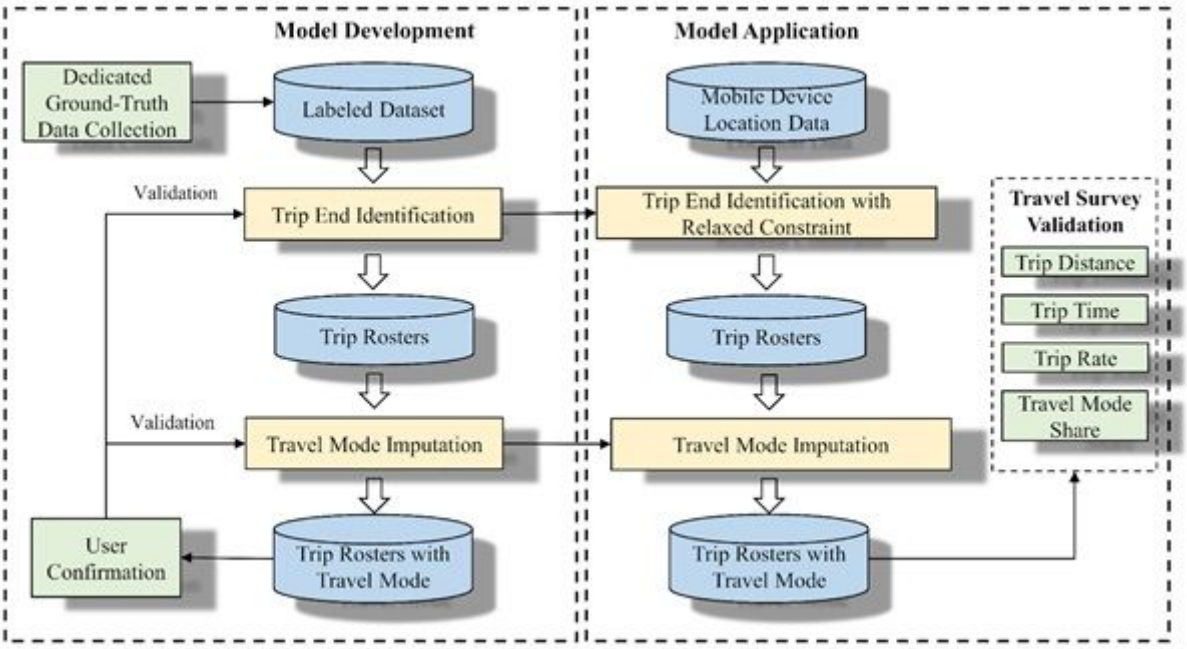

Figure 1

The Data-Driven Travel Mode Share Estimation Framework

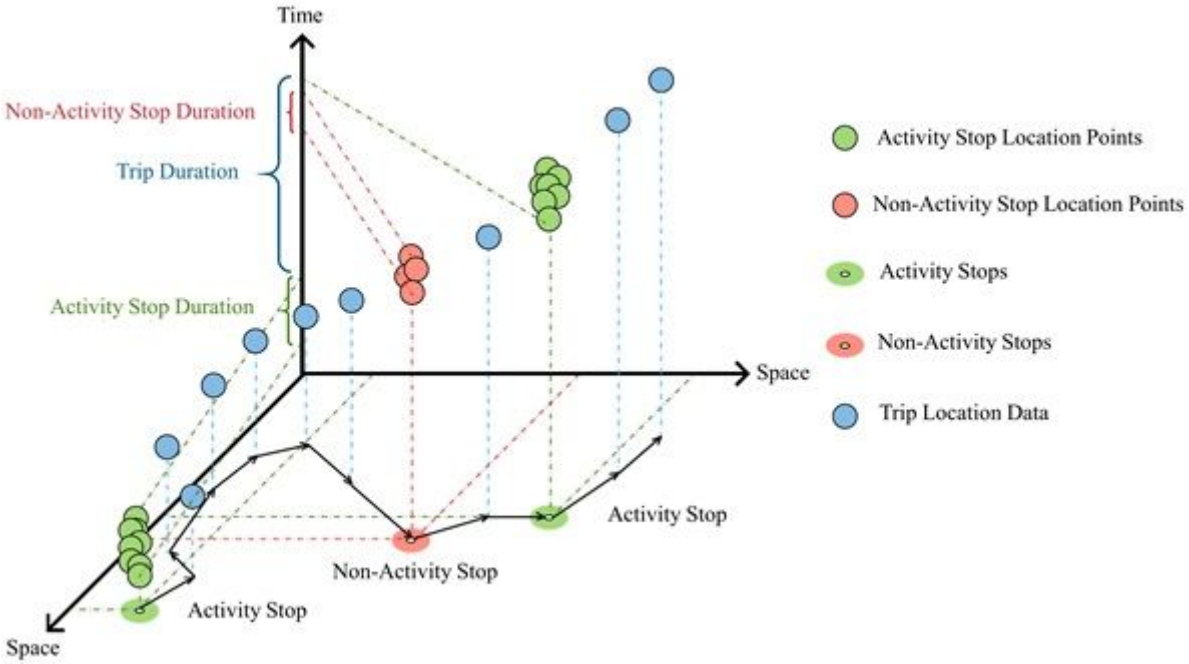

Figure 2

Typical Daily Travel Pattern of an Individual

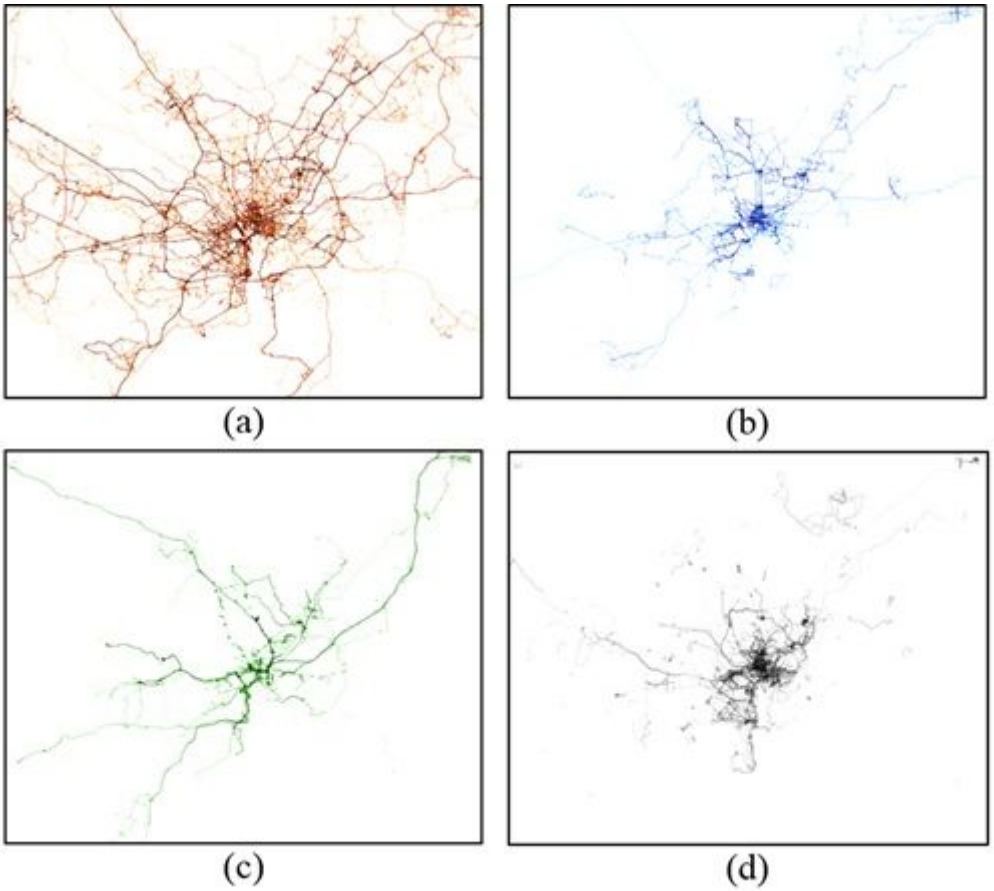

Figure 3

incenTrip Trip Trajectories for (a) drive; (b) bus; (c) rail; (d) non-motorized.

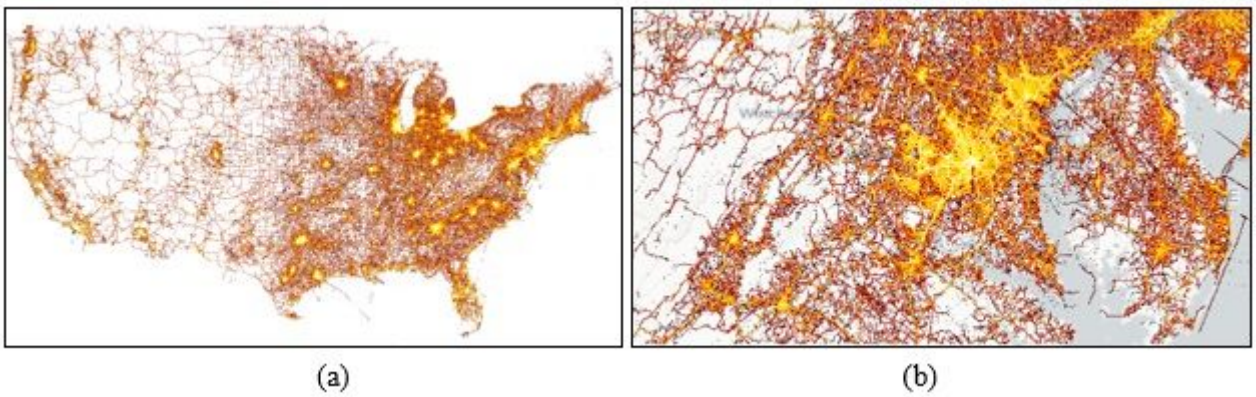

Figure 4

Raw Location Point Distribution for (a) Dataset I; (b) Dataset II.

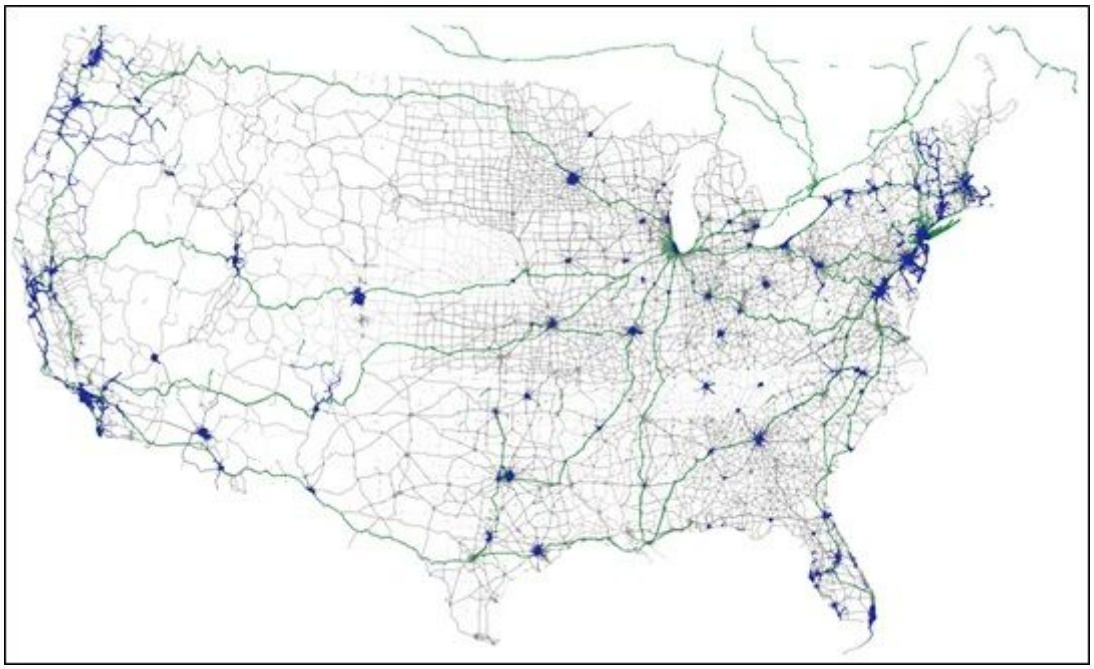

Figure 5

Multimodal Transportation Networks: drive (grey), rail (green), bus (blue).

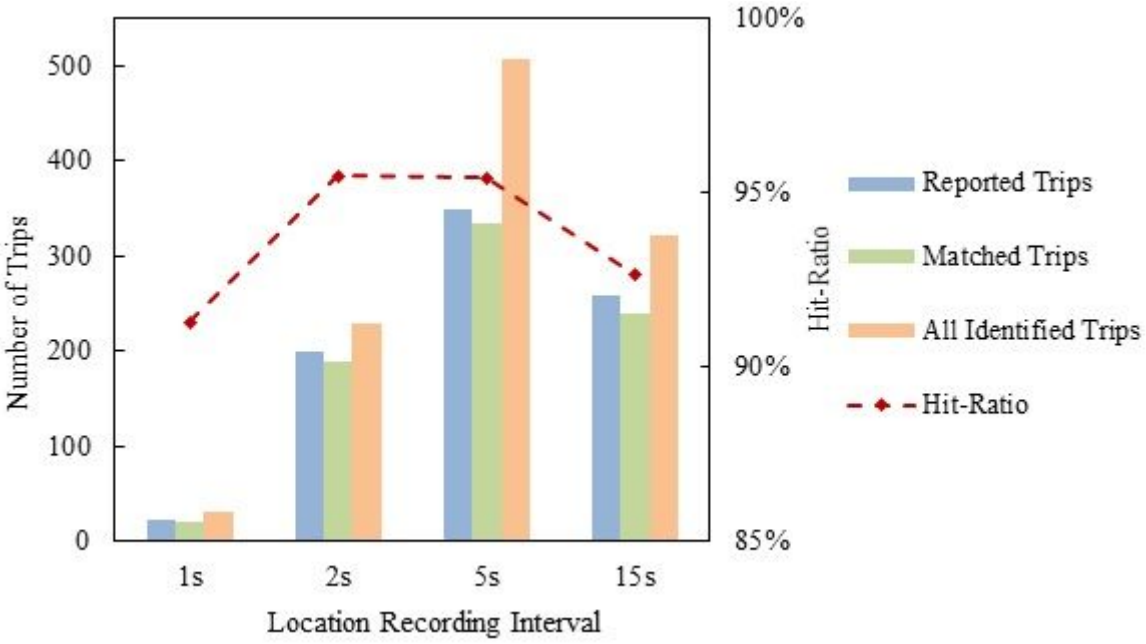

Figure 6

Trip End Identification Results.

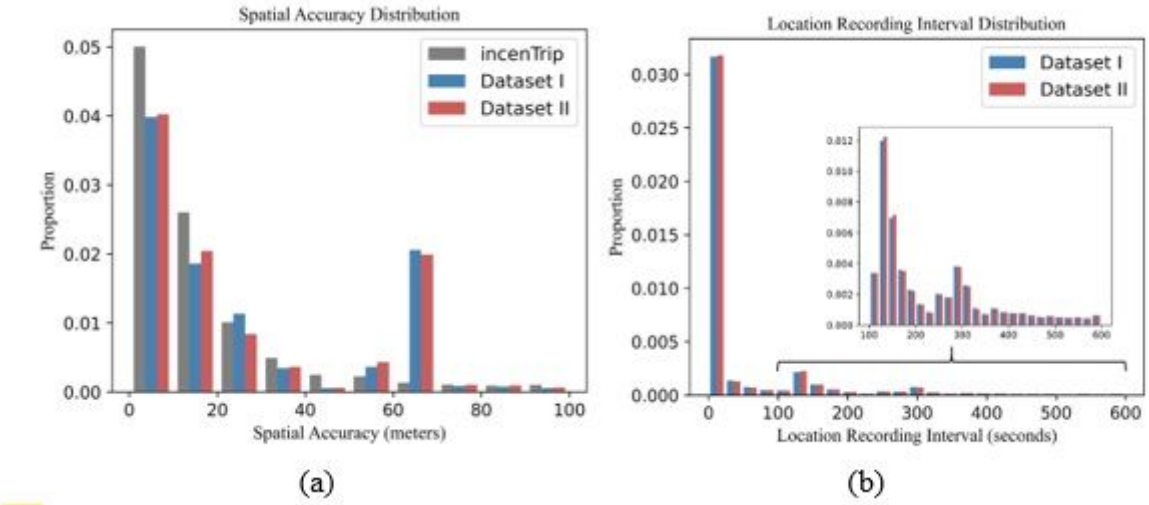

Figure 7

(a) Spatial accuracy distribution for the three LBS datasets; (b) Location recording interval distribution for the two case studies' LBS Datasets.

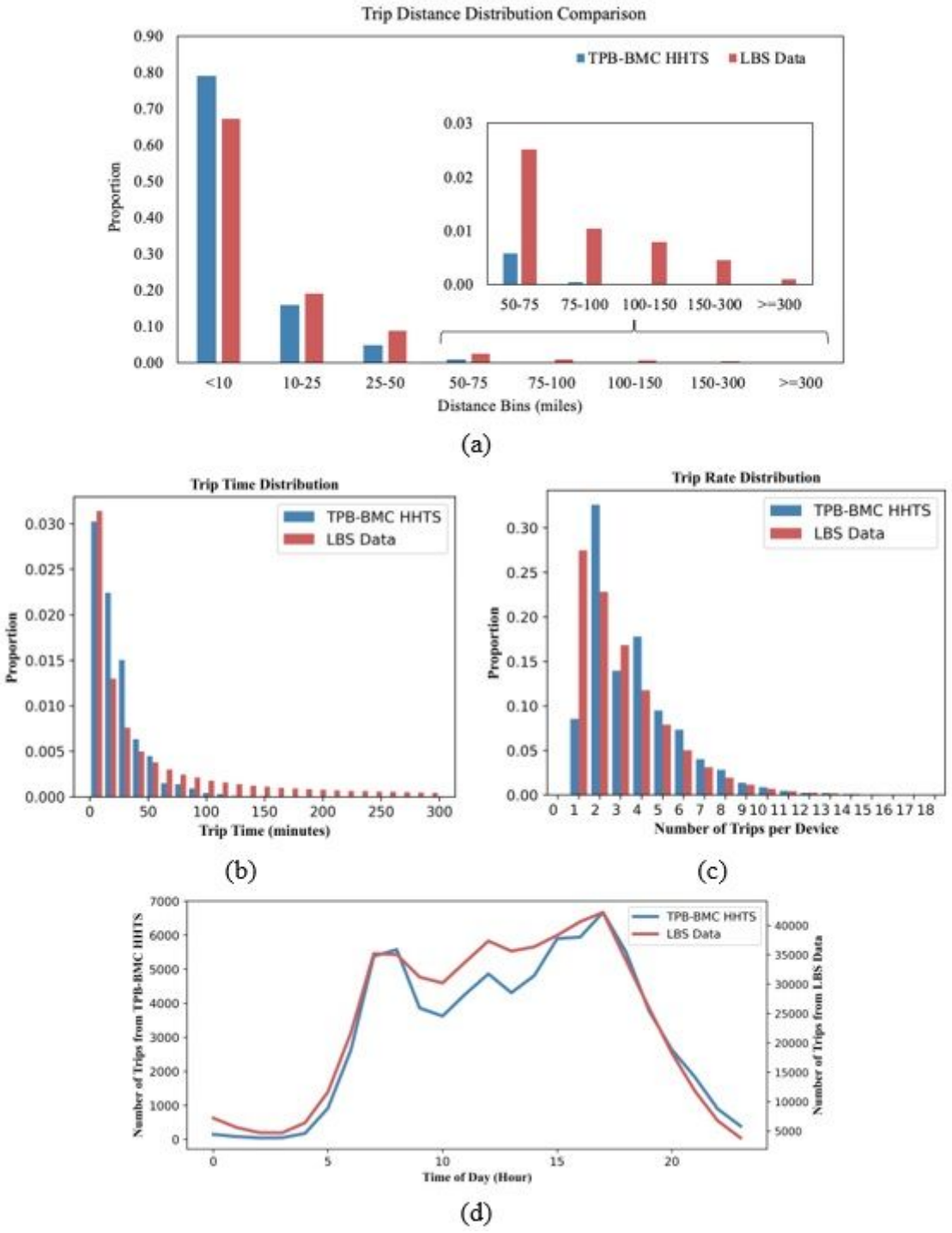

Figure 8

Comparison with 2007/08 TPB-BMC HHTS on: (a) Trip distance distribution; (b) Trip time distribution; (c) Trip rate distribution; (d) Time of day distribution.

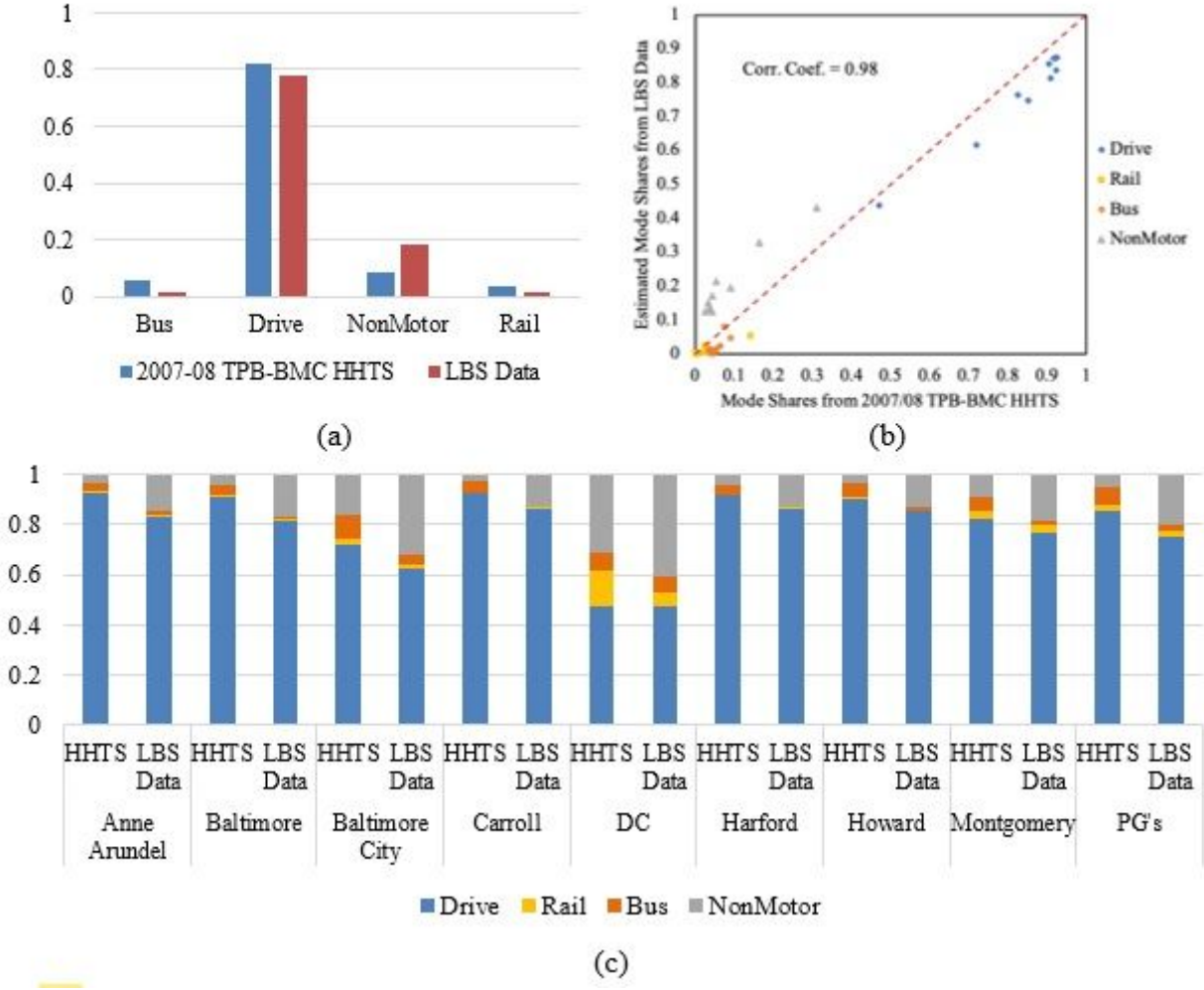

Figure 9

(a) Statewide travel mode share comparison; (b) County-level travel mode share correlation; (c) County-level travel mode share comparison.

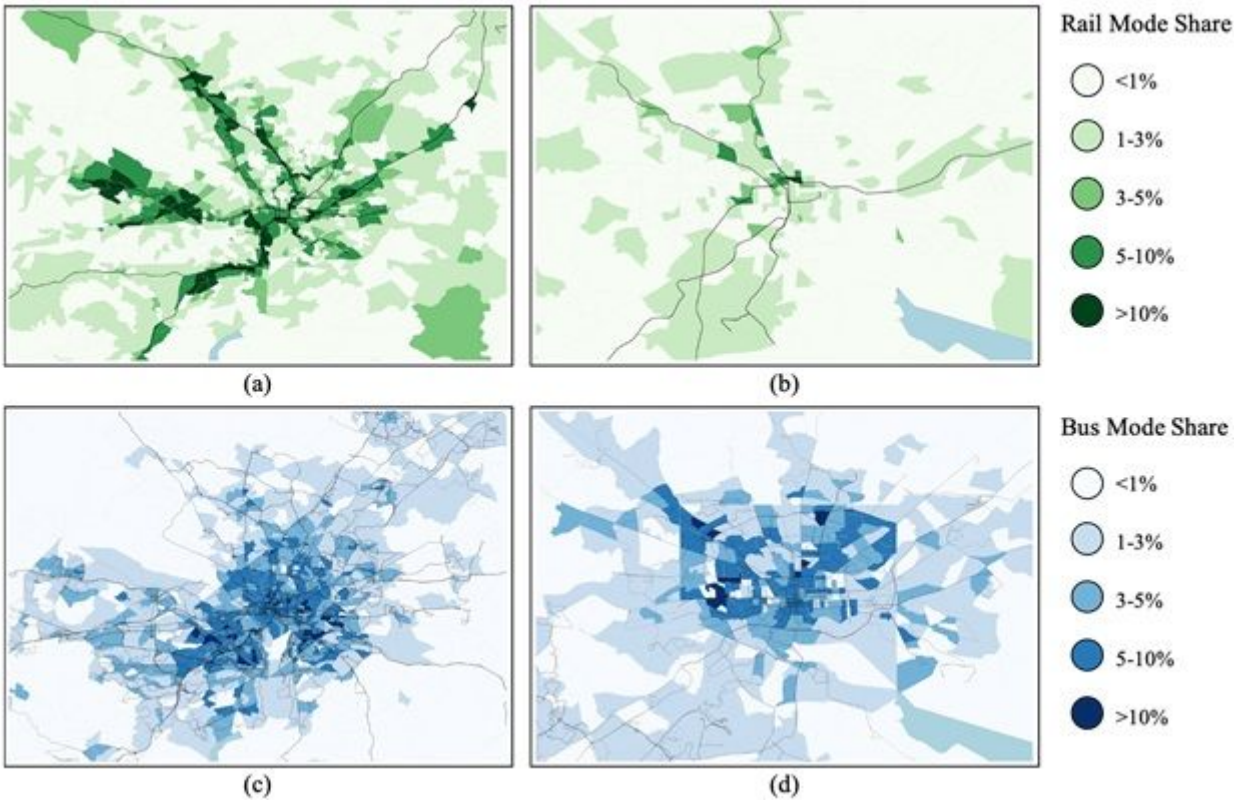

Figure 10

Census tract-level mode share illustration for (a) and (b): Rail mode share in D.C. and Baltimore City; (c) and (d): Bus mode share in D.C. and Baltimore City.

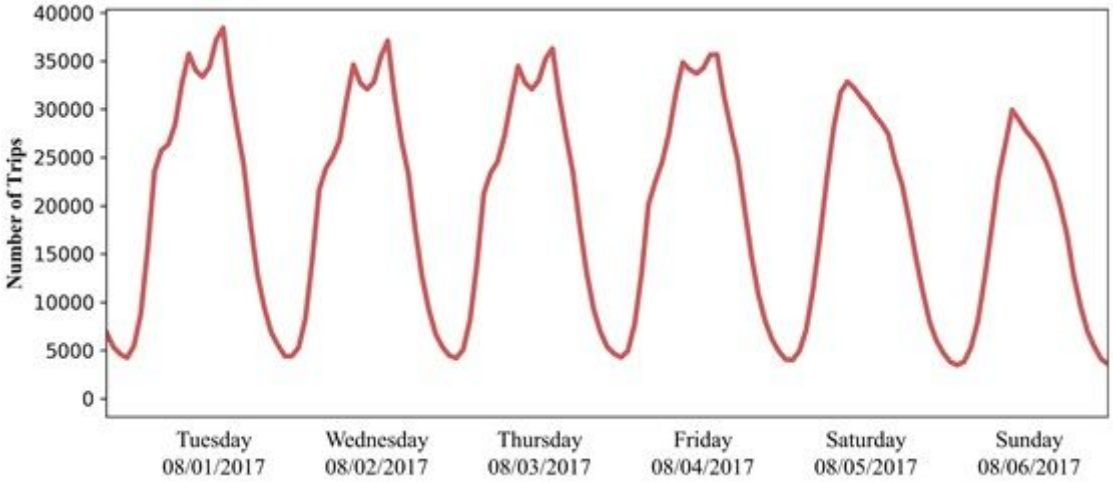

Figure 12

Day of Week Variation.

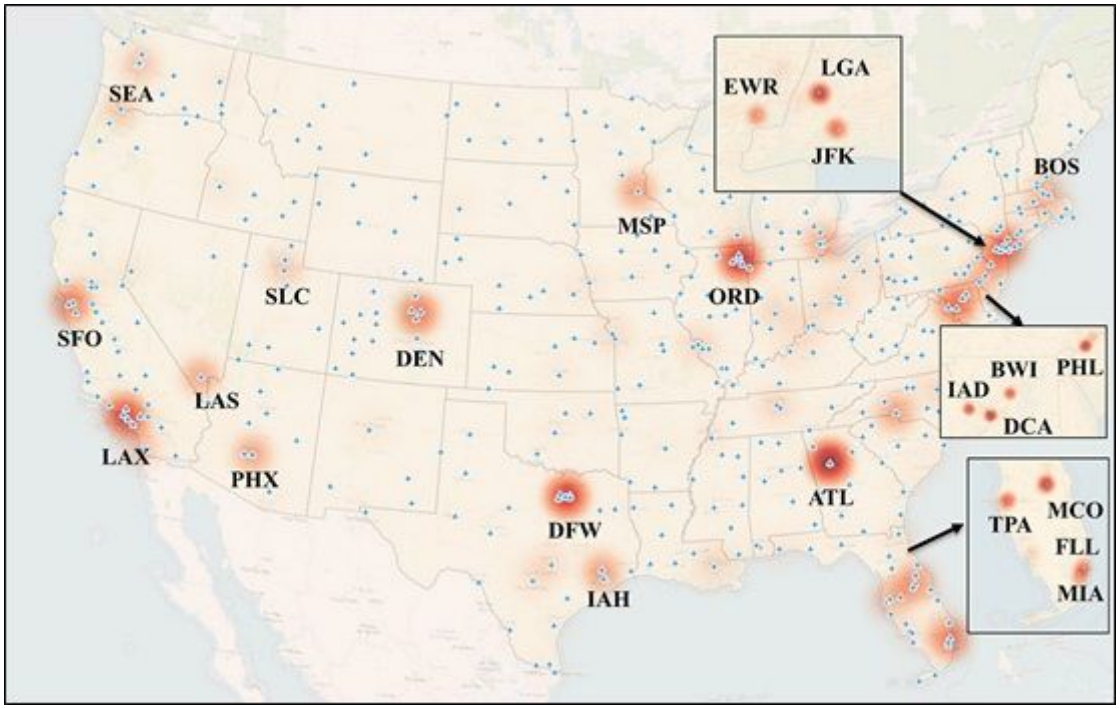

Figure 13

Nationwide-level Air Trip Origins

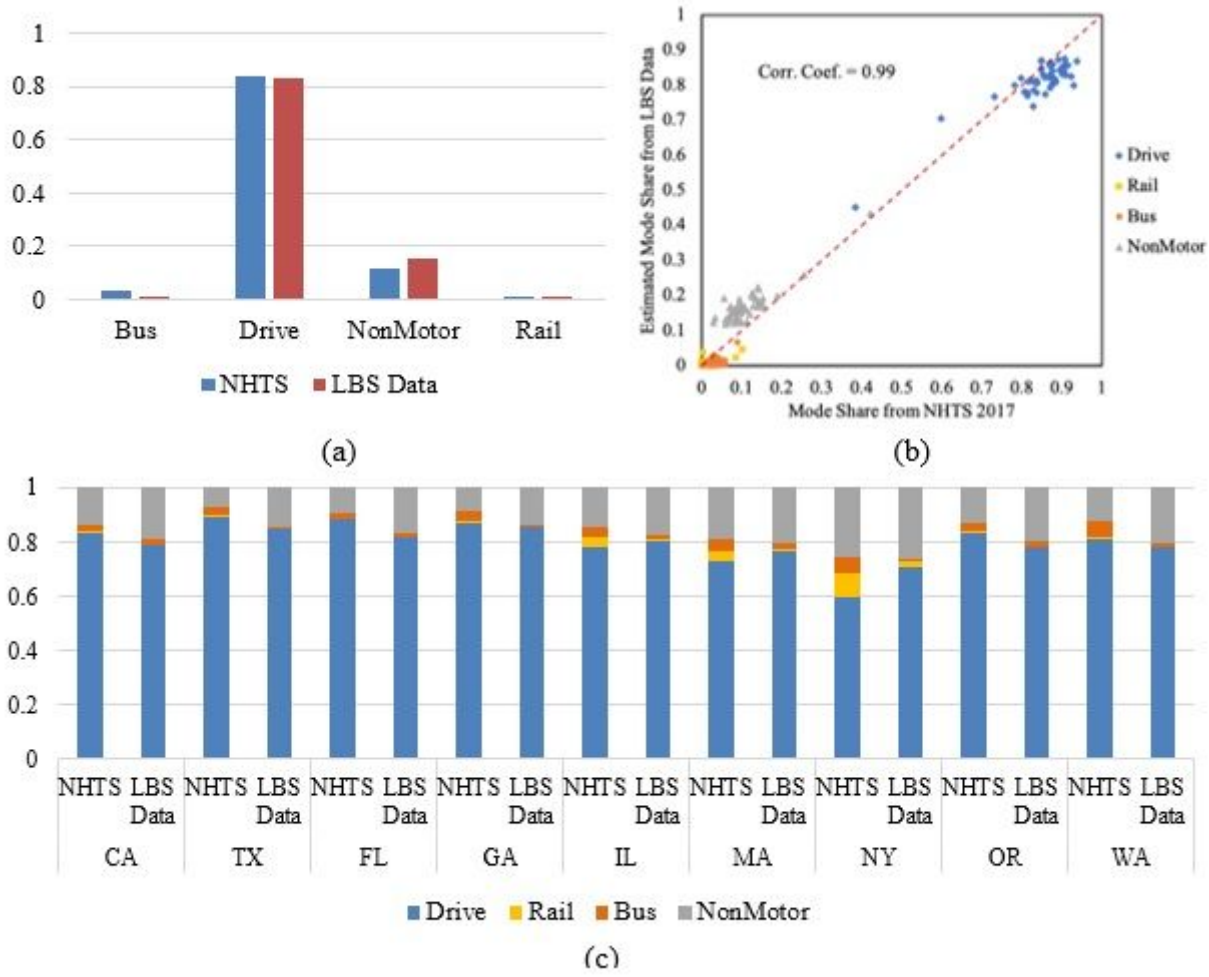

Figure 14

(a) Nationwide travel mode share comparison; (b) State-level travel mode share correlation; (c) State-level travel mode share comparison.

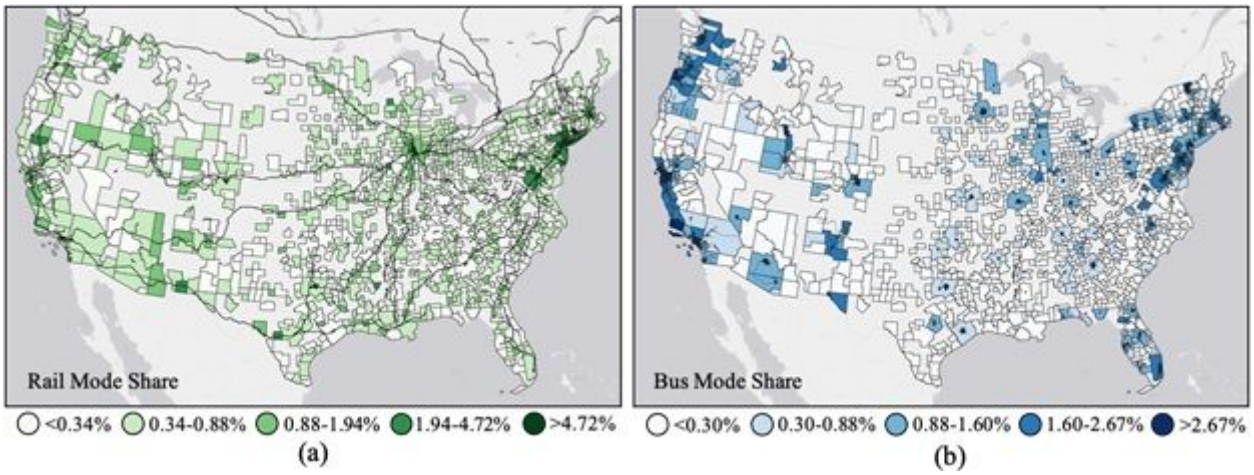

Figure 15

CBSA-level illustration of (a) rail travel mode share; (b) bus travel mode share.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendices.pdf