A data-driven travel mode share estimation framework based on mobile device location data

Mofeng Yang¹ · Yixuan Pan¹ · Aref Darzi¹ · Sepehr Ghader¹ ·
Chenfeng Xiong^{1,2} · Lei Zhang¹

Accepted: 13 July 2021 / Published online: 12 August 2021

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

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 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 over 93% tenfold 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.

Keywords Travel mode share · Travel surveys · Machine learning · Mobile device location data

✉ Lei Zhang
lei@umd.edu

Extended author information available on the last page of the article

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. “[Literature review](#)” section reviews the literature on MDLD. “[The data-driven travel mode share estimation framework](#)” section describes the proposed data-driven framework and models in detail. “[Data](#)” section introduces the data. “[Model development](#)” section presents the model development results. “[Case studies](#)” section demonstrates the framework with two large-scale case studies. “[Conclusions and discussions](#)” section concludes this paper and discusses the limitations and future research directions.

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.

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, 2004; Ojah and Pearson 2008; Wolf and Lee 2008; ETC Institute 2009, 2011a, b, c). 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, 2015). Some travel surveys utilized both in-vehicle and wearable GPS data loggers to take advantage of both technologies (NuStats 2007, 2011, 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 (Haghani 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. 2016). 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. 2016). 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. 2016). 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. 2012a, b; Pappalardo, et al. 2015; Song et al., 2010a, b; 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. 2010). 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 (Landmark 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. 2016; 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. 2016; Gonzalez et al. 2008; Kang et al. 2012a, b; 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, 2015a, b, 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, b).

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 m) and the lowest LRI (usually 1 s), 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.

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 and Shalaby 2006; McGowen and McNally 2007; Du and Aultman-Hall 2007; Stopher et al. 2008; Schuessler and Axhausen 2009; Bohte and Maat 2009; Gong et al. 2012; Gong et al. 2014; Safi et al. 2016; Patterson and Fitzsimmons 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, 2018; Zhou et al. 2016). 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 s, 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

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%
Brunnauer 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

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.

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

Fig. 1 The data-driven travel mode share estimation framework

before the data products are deemed useful and applicable for any transportation planning applications.

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_{acr} : maximum activity distance range. If the distance between two consecutive clusters stayed within s_{acr} , 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_{acr} :

Fig. 2 Typical daily travel pattern of an individual

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.

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 “Machine learning overview” section in Appendix 1.

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-m 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 m to calculate these features to obtain more accurate estimates. Accelerator meter data is also useful in travel mode

Table 2 Features constructed from mobile device location data

Features	Unit
<i>Location recording interval feature</i>	
Average # of records per minutes	Frequency
<i>Trip features</i>	
Trip distance	m
Origin–destination distance	m
Trip time	min
Max., Min., Avg, Med., 5-, 25-, 75-, 95-percentile speed	m/s
<i>Multimodal transportation network features (location data with accuracy < 50 m)</i>	
% of location points fell within 50 m of rail network	%
% of location points fell within 50 m of bus network	%
% of location points fell within 50 m of drive network	%
% of location points fell within 50 m of bus stops	%

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.

Data

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 *incentTrip* (*incentrip.org*). *incentTrip* 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). *incentTrip* 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 *incentTrip* data is set to be 1–15 s 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 *incentTrip* 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 *incentTrip* users were hired to travel with *incentTrip* 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

Fig. 3 incenTrip trip trajectories for **a** drive; **b** bus; **c** rail; **d** non-motorized

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

is randomly sampled and used for this study in order to reduce the computation burden, including 266,149 unique devices. Figure 4a, b visualizes the raw location points for Dataset I and Dataset II respectively, where the lighter area means more concentrated data.

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.

Fig. 4 Raw location point distribution for **a** Dataset I; **b** Dataset II

Travel surveys

Two travel surveys are used in this study for comparison purposes: NHTS 2017 and 2007/2008 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 (MWCOC 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.

Fig. 5 Multimodal transportation networks: drive (grey), rail (green), bus (blue). (Color figure online)

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

Fig. 6 Trip end identification results

Model development

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:

$$\begin{aligned}
 t &\geq n \cdot f \\
 t_{act} &\geq n \cdot f \\
 s_{act} &\geq s \\
 n \cdot f &\geq s/v
 \end{aligned}$$

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-m, 50-m, and 100-m as the candidate s value. Also, the t_{act} was set as 300 s 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” are the trips that are originally reported in the respondents’ travel diaries; the “Matched Trips” are the trips

identified from the data that match the trips in the “Reported Trips”; the “All Identified Trips” is the summation of the “Matched Trips” and the underreported trips that are not reported in respondents’ travel diaries but are identified by our algorithm and later confirmed by the respondents; and the “Hit-Ratio” (or recall rate) is calculated by dividing “Matched Trips” by “Reported Trips” to evaluate the performance of the algorithm. 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 the “Reported Trips” are identified for each LRI, with the overall Hit-Ratio equals to 94.5%. The algorithm also properly identifies the underreported trips, which consist of 15% to 35% of the total number of trips. Capturing these underreported trips help produce more detailed travel patterns for each respondent.

Travel mode imputation result

Five machine learning models, KNN, SVC, XGB, RF, and DNN are used to impute travel modes. A total of 6064 drive trips, 1824 rail trips, 1403 bus trips, 1496 bike trips, and 1901 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 “[Fine-tuned parameters for each machine learning method](#)” section in Appendix 1. During the model training process, tenfold cross-validation (CV) is used to fine-tune the model parameters with 70% training data and the tenfold CV accuracy is calculated by averaging each fold’s accuracy measurement, which is based on the correctly imputed trips for all modes. The fine-tuned models are then applied to the 30% testing data to calculate the F1 score for each travel mode. These measures are calculated using the equations as shown below:

$$\begin{aligned}
 10\text{-fold CV accuracy} &= \frac{1}{10} \sum_{i=1}^{10} \frac{\sum_k C_{k,i}}{T_i} \\
 Precision_k &= \frac{C_{k,test}}{P_{k,test}} \\
 Recall_k &= \frac{C_{k,test}}{R_{k,test}} \\
 F1_k &= 2 \cdot \frac{Precision_k \cdot Recall_k}{Precision_k + Recall_k}
 \end{aligned}$$

where $C_{k,i}$ represents the number of correctly imputed trips for travel mode k in fold i , T_i represents the total number of trips in fold i , $Precision_k$ is the precision measure for travel mode k , $C_{k,test}$ represents the number of correctly imputed trips for travel mode k in the testing data, $P_{k,test}$ represents the number of imputed trips for travel mode k in the testing data, $Recall_k$ is the recall measure for travel mode k , $R_{k,test}$ represents the actual number of trips for travel mode k in the testing data, $F1_k$ is the F1-score measure for travel mode k .

Table 5 compares the tenfold CV accuracy and the F1 scores for different models. RF achieved over 93% accuracy in both four and five modes models and performed the best among all 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

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
Tenfold 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
Tenfold CV accuracy	85.5	85.4	93.0	93.6	85.0

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

in Fig. 3a, 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 “[Feature importance of the random forest model](#)” section in Appendix 1.

Case studies

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 7a shows the spatial accuracy of the three datasets after removing the data with an accuracy greater than 100 m (see “[Spatial accuracy](#)” section in Appendix 2. for more details). The incen-Trip dataset has the best spatial accuracy among the three datasets, with more than 80% of the data's spatial accuracy less than 50 m, whereas the two LBS datasets show a bimodal distribution, with the second peak locates around 70 m. 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 m. Figure 7b 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

Fig. 7 **a** Spatial accuracy distribution for the three LBS datasets; **b** location recording interval distribution for the two case studies' LBS datasets

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 to 1800s; (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 “Spatial accuracy” and “Location recording interval” sections in Appendix 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 2007/2008 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 8a–c compare the distributions of trip distances, trip times, and trip rates estimated from the LBS data with the reported ones from the 2007/2008 TPB-BMC HHTS. A satisfactory match is observed between the two datasets, except that the LBS estimates show fewer shorter-distance trips and more longer-distance (separated by 50 miles threshold) and longer-duration trips (Hu and Reuscher 2004) (see more detailed trip distance comparisons in “Trip distance distribution comparisons” Appendix 3). The trip distance distribution estimated from the LBS data is directly calculated by summing up the distance between consecutive location points within a trip, while the survey distances are from the reported trip distances. In this case, the use of different distance calculation methods can

Fig. 8 Comparison with 2007/2008 TPB-BMC HHTS on: **a** trip distance distribution; **b** trip time distribution; **c** trip rate distribution; **d** time of day distribution

contribute to observing fewer shorter-distance trips in the LBS estimates. The higher proportion of longer-distance trips and longer-duration trips might also be attributed to the differences in sample size and spatiotemporal coverages between the survey and LBS data. It can be observed that, for 2007/2008 TPB-BMC HHTS, there is a sudden change between 50–75 miles and 75–100 miles. This might be because the 2007/2008 TPB-BMC HHTS only collected one-day travel diary from less than 14,000 households in the region.

Although the respondents tended not to underreport their long-distance trips, the limited sample size might still lead to this sudden change in addition to the overall underestimation of the regional long-distance travel demand. In addition, the LBS data sample is more likely to include the active travelers that make more trips and/or longer-duration trips, such as long-distance travel for leisure or business purposes or long-distance commute. The higher proportion of devices with only one-trip in the LBS estimates might be due to the technology limitation. As discussed in the previous section, the LBS data has high LRI and its variability, and the observed duration of each device varies. Therefore, for some devices, only a part of their daily trips can be captured in the LBS data, while the travel survey collects on the residents' all trips on a typical travel day (see detailed discussions on the observed duration in “Observed duration” section in Appendix 2). In summary, all these discrepancies are expected when comparing the LBS estimates with survey estimates because of the nature of the two data collection methods. The number of trips by the time of day is also compared using average weekday trips from the survey, as shown in Fig. 8d. It can be seen that both morning and afternoon peaks are matched in the two data sources, while LBS data can capture more daytime trips.

Figure 9a shows the statewide travel mode share comparison result. A satisfactory match can be achieved compared with the 2007/2008 TPB-BMC HHTS mode share, while lower bus and rail mode shares and higher non-motorized mode share are shown in the LBS estimates. One explanation is that in the travel survey, people tend to underreport

Fig. 9 a Statewide travel mode share comparison; b county-level travel mode share correlation; c county-level travel mode share comparison

short-distance walking trips. Also, signal loss and urban canyon effect also increase the difficulties in identifying and imputing the rail and bus trips, especially with the high-LRI LBS data. The travel mode share is further compared at eight counties in Maryland and D.C., as shown in Fig. 9b, c. Though a high Pearson correlation value (0.98) between the LBS estimates and the survey result is observed, the high correlation value largely depends on the high drive mode share and variations can also be seen for all four modes. However, it should be noted that the 2007/2008 TPB-BMC HHTS is conducted over ten years ago and the travel patterns have been significantly changed, especially for transit travels. According to BTS's latest Transportation Statistics Annual Report, the transit ridership in the U.S. surpassed 9 billion in 2007, reached a high of 9.9 billion trips in 2014, and declined in the subsequent years (U.S. DOT BTS 2020). Therefore, we compare the LBS estimates with the latest NHTS 2017. However, due to the small sample size of NHTS 2017 in some rural counties (e.g., Kent County, Charles County), their estimates are prone to be over/underestimated. A detailed comparison can be found in “[Travel mode share comparisons](#)” section in Appendix 3.

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.

Fig. 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

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. 11a–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 2007/2008 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.

Fig. 11 Comparison with NHTS 2017 on: **a** Trip distance distribution; **b** trip time distribution; **c** trip rate distribution

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 h, 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 h 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 14a shows the nationwide travel mode share comparison result. Still, a satisfactory match can be achieved after applying the travel mode imputation model to the national dataset. The differences observed between the nationwide LBS estimates and NHTS 2017 are even smaller than the statewide comparison with 2007/2008 TPB-BMC HHTS, resulting in a higher Pearson correlation value (0.99) (Fig. 15b) across 50 states and D.C. One possible reason for this smaller difference can be attributed to the data itself, since the LBS data used in this study is also collected in 2017, which is around the time when NHTS 2017 was conducted. These results demonstrate the transferability and the generalization ability of the proposed framework that can be applied in different and larger geographies. We also selected eight states across the U.S. for detailed comparisons, as shown in Fig. 14c. Though the general pattern is captured, an overall underestimation of bus and rail trips can still be observed, especially for IL, MA, and NY, where three major cities (Chicago, IL; Boston, MA; New York City, NY) that have large rail and bus services located in. This can also be attributed to more serious signal loss and urban canyon effect because of the large number of skyscrapers. Detailed comparison for all fifty states and D.C. can be found in “Travel mode share comparisons” section in Appendix 3.

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

Fig. 12 Day of week variation

Fig. 13 Nationwide-level air trip origins

Fig. 14 a Nationwide travel mode share comparison; b state-level travel mode share correlation; c state-level travel mode share comparison

Fig. 15 CBSA-level illustration of **a** rail travel mode share; **b** bus travel mode share

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.

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% tenfold 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, 2007/2008 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, b; 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.

Appendix 1: Travel mode imputation with machine learning models

Machine learning overview

KNN is one of the earliest and simplest classification models (Peterson 2009). The main idea of KNN is to find the top k nearest samples of the target sample based on distance measurement. The distance between two samples is usually calculated based on Euclidean distance. SVC was developed by Cortes and Vapnik in the 1990s and applied for face recognition, pattern recognition, etc. (Cortes and Vapnik 1995; Osuna et al. 1997; Suykens and Vandewalle 1999; Wang 2005). 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), sigmoid, etc. XGB is one of the most recent ensemble-learning algorithms using the boosting technique (Chen et al. 2015; Chen and Guestrin 2016). 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 (Kearns and Valiant 1988; Kearns and Valiant 1994). RF is one of the most famous ensemble-learning algorithms using the bagging technique (Liaw and Wiener 2002; Breiman 1996). Bagging (Bootstrap aggregating) is a machine learning technique that tends to improve the stability and accuracy by generating multiple training sets by sampling from the data uniformly and with replacement (Breiman 1996). 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 (Liaw and Wiener 2002). In short, RF is essentially a collection of decision trees (Quinlan 1986) 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. DNN is an Artificial Neural Network (ANN) with multiple layers between the input and output layers (Bengio 2009; Schmidhuber 2015). 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 (Hecht-Nielsen 1992). Detailed methodology of DNN and ANN can be found in Bengio (2009) and Schmidhuber (2015).

Fine-tuned parameters for each machine learning method

See Table 6.

Table 6 Fine-tuned hyperparameters for different machine learning models

Four travel modes		Five travel modes	
Model	Parameters	Model	Parameters
KNN	k = 5; p = 1; weights = 'distance'; algorithm = 'brute'; leaf_size = 1; metric = 'minkoski'	KNN	k = 1; p = 1; weights = 'distance'; algorithm = 'ball_tree'; leaf_size = 1; metric = 'minkoski'
SVC	Cs = 100; Gammas = 1; class_weight = 'None'; kernel = 'rbf'	SVC	Cs = 100; Gammas = 1; class_weight = 'None'; kernel = 'rbf'
RF	n_estimators = 700; max_features = 'sqrt'; max_depth = 30; min_samples_split = 2; min_samples_leaf = 1; bootstrap = 'False'; class_weight = 'None'	RF	n_estimators = 800; max_features = 'auto'; max_depth = 60; min_samples_split = 2; min_samples_leaf = 1; bootstrap = 'False'; class_weight = 'balanced_subsample'
XGB	colsample_bytree = 0.980; gamma = 0.429; learning_rate = 0.317; max_depth = 5; n_estimators = 136; subsample = 0.627	XGB	colsample_bytree = 0.865; gamma = 0.299; learning_rate = 0.324; max_depth = 5; n_estimators = 126; subsample = 0.884
DNN	learning_rate = 0.001; epochs = 500; batch_size = 32; optimizer: rmsprop; loss: 'categorical_crossentropy'; layers: 64–32–16–8–DropOut (0.2)	DNN	learning_rate = 0.001; epochs = 500; batch_size = 32; optimizer: rmsprop; loss: 'categorical_crossentropy'; layers: 64–32–16–8–DropOut (0.2)

Feature importance of the random forest model

Figure 16 shows the feature importance value of the RF model for four and five travel modes respectively. The feature importance value (Gini importance) is automatically calculated using the python package scikit-learn (Pedregosa et al. 2011), 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. In addition, the percentage of records that fell within 50 m of the rail network is also significantly important since it is a representative feature to impute rail trips.

Fig. 16 Feature importance of the random forest classifier

Appendix 2: Spatial accuracy, location recording interval, and observed duration of the location-based service data

Spatial accuracy

Table 7 shows the detailed spatial accuracy distribution of the three datasets. The spatial accuracy information is one of the optional features in the most commercially available LBS data sources. The metric is usually derived from various spatial positioning techniques and provides the horizontal accuracy of each latitude and longitude record reported in the dataset with a 95% confidence interval in most applications. In this study, the LBS data sources all have that optional feature, and further analysis on the spatial accuracy of each observation can be conducted relying on that information. It can be clearly seen that the incenTrip dataset has the highest data quality among the three datasets, with over 80% of the data has a spatial accuracy smaller than 50-m. This small location uncertainty could also be achieved through positioning techniques like Wi-Fi and Bluetooth in addition to GPS. For the other two large-scale LBS datasets, the spatial accuracy shows a similar bimodal distribution, with the second peak locates around 70-m. In general, for all three datasets, around 90% of the data has spatial accuracy smaller than 100-m. Therefore, before we apply our proposed framework to the dataset, we removed the location data with spatial accuracy larger than 100-m for all three datasets to ensure the data quality.

Table 7 Spatial accuracy distribution of the three datasets

Spatial accuracy	incenTrip		Dataset I		Dataset II	
	Proportion (%)	Cumulation (%)	Proportion (%)	Cumulation (%)	Proportion (%)	Cumulation (%)
(0, 10]	44.15	44.15	49.91	49.91	50.21	50.21
(10, 20]	22.50	66.64	7.58	57.49	5.78	55.99
(20, 30]	8.79	75.43	6.80	64.28	7.27	63.27
(30, 40]	4.29	79.73	1.00	65.28	1.07	64.33
(40, 50]	2.17	81.90	3.47	68.76	4.10	68.43
(50, 60]	2.01	83.92	0.26	69.01	0.28	68.71
(60, 70]	1.10	85.01	18.33	87.34	17.60	86.31
(70, 80]	0.92	85.93	0.69	88.04	0.83	87.14
(80, 90]	0.79	86.73	0.66	88.70	0.80	87.94
(90, 100]	1.19	87.92	0.59	89.29	0.66	88.60
(100, 200]	3.81	91.73	4.18	93.46	4.87	93.47
(200, 500]	2.57	94.31	0.30	93.77	0.23	93.70
(500, ...]	5.69	100.00	6.23	100.00	6.30	100.00

Location recording interval

Table 8 shows the detailed LRI distributions for the two large-scale LBS datasets after removing the location data with spatial accuracy larger than 100-m, which also follows the bimodal distribution, with the second peak locates around 120-s. In both datasets, more than 50% of the data has an LRI smaller than 30 s.

Table 8 Location recording frequency distribution of the two case studies' datasets

LRI	Dataset I		Dataset II	
	Proportion (%)	Cumulation (%)	Proportion (%)	Cumulation (%)
(0, 10]	44.81	44.81	46.48	46.48
(10, 20]	4.98	49.79	4.92	51.40
(20, 30]	1.83	51.62	1.77	53.17
(30, 40]	1.21	52.83	1.15	54.32
(40, 50]	0.86	53.69	0.81	55.13
(50, 60]	0.73	54.42	0.68	55.81
(60, 70]	0.64	55.06	0.60	56.42
(70, 80]	0.50	55.56	0.48	56.90
(80, 90]	0.45	56.01	0.43	57.32
(90, 100]	0.38	56.39	0.36	57.69
(100, 150]	6.46	62.85	6.70	64.39
(150, 200]	2.78	65.63	2.75	67.14
(200, 250]	1.11	66.73	1.08	68.22
(250, 300]	2.11	68.84	2.13	70.34
(300, 350]	1.33	70.17	1.30	71.64
(350, 400]	0.75	70.92	0.75	72.40
(400, 500]	1.06	71.98	1.06	73.46
(500, ...]	28.02	100.00	26.54	100.00

Observed duration

The observed duration distribution is important information that examines how much information is covered temporally for each device. Due to technology limitation, the LBS can only capture a subset of device’s movements based on the mobile device usages and might not be able to cover a full trip for the device, which may lead to biases and affect the reliability of the LBS estimates. Figure 17 and Table 9 show the observed duration distributions of Dataset I and Dataset II. It can be seen that: (1) for Dataset I, around 60% of the devices are observed at least 8 h and more than 80% of the devices are observed more than 4 h; (2) for Dataset II, since it covers one week of data for the entire U.S., the unique hour bins are used to measure its coverage. Similarly, more than 80% of the devices can be observed for at least 29–35 h (4–5 h per day) in a week. In addition, more than 70% of the devices are observed at least 5 days in a week. Though the observed duration distributions show that a large percentage of the devices have relatively consistent observations along the time, still, there are devices that are less observed.

Fig. 17 Observed duration distribution for: **a** number of unique hours (Dataset I); **b** unique hour bins (Dataset II); (3) number of unique days (Dataset II)

Table 9 Observed duration distribution of the two case studies' datasets

Number of unique hours	Dataset I		Unique hour bins	Dataset II	
	Proportion (%)	Cumulation (%)		Proportion (%)	Cumulation (%)
24	3.12	3.12	162–168	0.99	0.99
23	1.93	5.05	155–161	1.36	2.35
22	1.87	6.92	148–154	1.60	3.95
21	2.05	8.97	141–147	1.83	5.78
20	2.20	11.17	134–140	2.08	7.86
19	2.43	13.60	127–133	2.31	10.17
18	2.64	16.24	120–126	2.61	12.78
17	2.95	19.19	113–119	2.93	15.71
16	3.24	22.43	106–112	3.31	19.02
15	3.56	25.99	99–105	3.79	22.81
14	3.91	29.90	92–98	4.35	27.16
13	4.23	34.13	85–91	5.00	32.16
12	4.56	38.69	78–84	5.65	37.81
11	4.85	43.54	71–77	6.36	44.17
10	5.08	48.62	64–70	6.84	51.01
9	5.35	53.97	57–63	7.03	58.04
8	5.37	59.34	50–56	7.01	65.05
7	5.49	64.83	43–49	6.60	71.65
6	5.57	70.40	36–42	5.97	77.62
5	5.56	75.96	29–35	5.19	82.81
4	5.63	81.59	22–28	4.44	87.25
3	5.51	87.10	15–21	3.98	91.23
2	5.98	93.08	8–14	3.80	95.03
1	6.92	100.00	1–7	4.97	100.00

Appendix 3: Detailed comparison results

Trip distance distribution comparisons

See Figs. 18 and 19.

Fig. 18 Comparison with 2007/2008 TPB-BMC HHTS on: **a** short-distance trip distance distribution; **b** long-distance trip distance distribution

Fig. 19 Comparison with NHTS 2017 on: **a** short-distance trip distance distribution; **b** long-distance trip distance distribution

Travel mode share comparisons

See Tables 10, 11, 12 and Figs. 20 and 21.

Table 10 County-level travel mode share comparison between Dataset I and 2007/2008 TPB-BMC HHTS

County	2007–2008 TPB-BMC HHTS				lbs data			
	Drive (%)	Rail (%)	Bus (%)	NonMotor (%)	Drive (%)	Rail (%)	Bus (%)	NonMotor (%)
Anne Arundel	92.285	0.728	3.666	3.321	83.351	0.694	1.030	14.925
Baltimore	90.773	0.741	4.106	4.379	81.234	0.628	1.324	16.814
Baltimore City	71.962	2.555	9.163	16.320	62.658	1.140	4.383	31.819
Carroll	92.411	0.361	4.705	2.523	86.340	0.704	0.050	12.906
DC	47.160	14.225	7.533	31.083	47.129	5.532	6.568	40.771
Harford	91.646	0.115	3.891	4.348	86.221	0.876	0.253	12.650
Howard	90.461	0.464	5.294	3.780	85.274	0.441	1.096	13.189
Montgomery	82.504	2.719	5.749	9.028	76.979	2.917	1.446	18.659
Prince George's	85.244	2.856	6.474	5.425	75.069	2.187	2.255	20.489

Table 11 County-level travel mode share comparison between Dataset I with NHTS 2017

County	2007–2008 TPB-BMC HHTS				lbs data			
	Drive (%)	Rail (%)	Bus (%)	NonMotor (%)	Drive (%)	Rail (%)	Bus (%)	NonMotor (%)
Allegany	87.807	0.000	2.959	9.234	80.437	1.405	0.104	18.054
Anne Arundel	87.724	0.016	2.235	10.024	83.351	0.694	1.030	14.925
Baltimore	85.655	0.403	5.386	8.556	81.234	0.628	1.324	16.814
Baltimore City	71.348	0.095	10.556	18.001	62.658	1.140	4.383	31.819
Calvert	93.816	0.000	2.456	3.728	87.470	0.747	0.222	11.562
Caroline	83.464	0.000	9.353	7.183	85.516	0.781	0.087	13.617
Carroll	89.947	0.000	10.053	0.000	86.340	0.704	0.050	12.906
Cecil	85.739	0.000	2.311	11.950	86.746	0.636	0.021	12.598
Charles	76.195	0.000	22.494	1.311	86.656	0.510	0.382	12.452
DC	94.103	0.000	1.249	4.648	47.129	5.532	6.568	40.771
Dorchester	89.483	0.218	3.028	7.271	83.926	0.464	0.000	15.611
Frederick	96.711	0.000	1.845	1.444	85.190	0.750	0.123	13.938
Garrett	83.215	0.000	11.136	5.649	87.016	0.749	0.000	12.235
Harford	88.492	0.296	3.917	7.295	86.221	0.876	0.254	12.650
Howard	53.718	0.000	0.000	46.282	85.274	0.441	1.096	13.189
Kent	82.855	2.218	2.666	12.261	80.349	0.465	0.000	19.186
Montgomery	82.108	3.363	6.060	8.469	76.979	2.917	1.446	18.659
Prince George's	86.707	0.088	9.284	3.921	75.069	2.187	2.255	20.489
Queen Anne's	79.223	0.000	13.979	6.798	88.069	0.558	0.147	11.225
Somerset	94.433	0.000	0.000	5.567	78.926	0.403	0.134	20.537
St. Mary's	91.749	0.070	4.866	3.315	85.908	0.718	0.112	13.262
Talbot	83.786	0.098	4.813	11.303	86.770	0.495	0.106	12.628
Washington	88.606	0.000	4.395	7.000	84.030	0.506	0.000	15.465
Wicomico	96.828	0.000	0.209	2.962	82.988	0.348	0.070	16.594
Worcester	38.446	10.277	9.038	42.239	78.947	0.301	0.030	20.722

Table 12 State-level travel mode share comparison between Dataset II and NHTS 2017

State	NHTS 2017				LBS Data			
	Drive (%)	Rail (%)	Bus (%)	NonMotor (%)	Drive (%)	Rail (%)	Bus (%)	NonMotor (%)
AL	93.730	0.000	3.228	3.042	86.991	0.365	0.257	12.387
AK	85.807	0.000	1.555	12.638	77.536	1.034	0.447	20.984
AZ	86.671	0.261	2.687	10.381	82.416	0.458	0.856	16.270
AR	90.110	0.000	3.152	6.737	85.397	0.367	0.246	13.990
CA	82.996	0.877	2.391	13.736	78.516	0.696	1.765	19.024
CO	83.664	0.291	2.051	13.994	81.318	0.637	0.778	17.267
CT	87.004	0.270	3.607	9.119	79.353	3.968	1.229	15.449
DE	85.623	0.000	3.052	11.325	82.412	0.421	1.244	15.923
DC	38.446	10.277	9.038	42.239	45.271	4.620	6.551	43.558
FL	88.334	0.139	2.252	9.276	81.840	0.352	1.035	16.772
GA	87.039	0.432	3.891	8.637	85.424	0.421	0.536	13.619
HI	82.758	0.000	3.081	14.162	74.049	0.318	3.007	22.625
ID	83.426	0.000	2.374	14.199	80.734	0.418	0.533	18.315
IL	78.180	3.295	4.064	14.460	80.112	1.216	1.021	17.650
IN	91.515	0.027	2.421	6.037	85.844	0.472	0.302	13.382
IO	88.007	0.000	2.712	9.280	81.602	0.415	0.202	17.782
KS	90.171	0.000	1.872	7.957	83.279	0.400	0.250	16.070
KY	87.221	0.000	3.837	8.942	86.451	0.293	0.659	12.597
LA	84.754	0.009	3.173	12.064	84.861	0.437	0.042	14.660
ME	84.901	0.000	3.924	11.175	86.913	0.352	0.171	12.563
MD	83.926	0.904	5.396	9.774	80.768	0.868	1.061	17.303
MA	73.094	3.804	4.156	18.945	76.801	0.932	2.183	20.084
MI	86.918	0.066	3.434	9.582	86.202	0.410	0.245	13.143
MN	86.800	0.528	3.461	9.211	82.265	0.460	0.704	16.571
MS	90.762	0.000	3.260	5.978	87.509	0.415	0.060	12.016
MO	87.743	0.048	3.744	8.465	84.161	0.408	0.562	14.868
MT	87.424	0.000	3.060	9.516	80.275	0.424	0.180	19.122
NE	87.648	0.000	1.800	10.551	82.419	0.387	0.144	17.050
NV	81.382	0.306	4.043	14.269	76.935	0.230	0.954	21.881
NH	84.751	0.000	5.688	9.561	87.187	0.333	0.299	12.181
NJ	80.564	1.541	2.591	15.303	78.124	1.221	1.324	19.330
NM	88.419	0.109	2.199	9.273	80.942	0.518	1.008	17.531
NY	59.810	8.620	5.983	25.586	70.472	2.194	1.342	25.992
NC	89.009	0.063	3.037	7.891	87.275	0.361	0.281	12.084
ND	92.371	0.000	0.419	7.210	82.452	0.465	0.050	17.033
OH	87.748	0.018	3.163	9.071	85.945	0.296	0.439	13.320
OK	90.560	0.000	1.462	7.978	84.098	0.294	0.284	15.324
OR	83.564	0.604	2.745	13.087	77.947	0.538	1.964	19.551
PA	81.608	0.831	4.727	12.833	81.196	0.628	1.176	16.999
RI	79.818	0.024	4.859	15.299	82.179	0.399	0.080	17.342
SC	91.041	0.026	2.392	6.542	85.958	0.307	0.224	13.511
SD	92.903	0.102	1.488	5.507	79.965	0.371	0.283	19.380
TN	90.954	0.077	2.716	6.254	86.290	0.322	0.438	12.950

Table 12 (continued)

State	NHTS 2017				LBS Data			
	Drive (%)	Rail (%)	Bus (%)	NonMotor (%)	Drive (%)	Rail (%)	Bus (%)	NonMotor (%)
TX	89.527	0.205	3.018	7.250	84.423	0.371	0.529	14.678
UT	87.304	0.529	1.275	10.892	82.325	0.528	1.039	16.108
VT	82.510	0.000	3.077	14.413	81.295	0.344	1.575	16.786
VA	85.252	1.169	3.716	9.863	83.307	0.887	0.714	15.093
WA	81.362	0.580	5.645	12.413	77.979	0.511	1.490	20.020
WV	90.802	0.000	5.759	3.439	85.569	0.473	0.139	13.819
WI	87.135	0.013	3.172	9.679	82.889	0.551	0.674	15.886
WY	90.873	0.000	2.204	6.923	83.101	0.263	0.020	16.616

Fig. 20 County-level travel mode share comparison between Dataset I and NHTS 2017

Fig. 21 State-level travel mode share comparison between Dataset II and NHTS 2017

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.

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.

Declaration

Conflict of interest The authors declare that they have no conflict of interest.

References

- 2000–2001 California Statewide Household Travel Survey. Final Report. NuStats, Austin (2002)
- 2010–2012 California Household Travel Survey. Final Report Version 1.0. NuStats, Austin (2013)
- 2010–2012 Minneapolis – St. Paul Travel Behavior Inventory. Twin Cities Metropolitan Council (2012)
- 2011 Atlanta, Georgia, Regional Travel Survey. Final Report. NuStats, Austin (2011)
- 2012–2013 Delaware Valley Household Travel Survey. Delaware Valley Regional Planning Commission (2013)
- 2014 Southern Nevada Household Travel Survey. Final Report. Westat, Rockville (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, DC (2003)
- Bachir, D., Khodabandelou, G., Gauthier, V., El Yacoubi, M., Puchinger, J.: Inferring dynamic origin-destination flows by transport mode using mobile phone data. *Transp. Res. Part C Emerg. Technol.* **101**, 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. *Found. Trends® Mach. Learn.* **2**(1), 1–127 (2009)
- Birant, D., Kut, A.: ST-DBSCAN: an algorithm for clustering spatial-temporal data. *Data Knowl. Eng.* **60**(1), 208–221 (2007)
- Bohte, W., Maat, K.: Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: a large-scale application in the Netherlands. *Transp. Res. C Emerg. Technol.* **17**(3), 285–297 (2009)
- Breiman, L.: Bagging predictors. *Mach. Learn.* **24**(2), 123–140 (1996)
- Breyer, N., Gundelgård, D., Rydergren, C.: Travel mode classification of intercity trips using cellular network data. *Transp. Res. Procedia* **52**, 211–218 (2021)
- Broach, J., Dill, J., McNeil, N.W.: Travel mode imputation using GPS and accelerometer data from a multi-day travel survey. *J. Transp. Geogr.* **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., 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. *Transp. Res. C Emerg. Technol.* **114**, 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. *Int. J. Online Eng.* **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, T., Guestrin, C.: Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining (2016)
- 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. *Transp. Res. C Emerg. Technol.* **68**, 285–299 (2016)
- Chicago Regional Household Travel Inventory: Draft Final Report. NuStats, Austin, Tex., and GeoStats, Atlanta (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)
- Çolak, S., Lima, A., González, M.C.: Understanding congested travel in urban areas. *Nat. Commun.* **7**, 10793 (2016)
- Cortes, C., Vapnik, V.: Support-vector networks. *Mach. Learn.* **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](https://arxiv.org/abs/1801.02143) (2018)
- Dabiri, S., Heaslip, K.: Inferring transportation modes from GPS trajectories using a convolutional neural network. *Transp. Res. C Emerg. Technol.* **86**, 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. *Transp. Res. A Policy Pract.* **41**(3), 220–232 (2007)
- Eagle, N., Macy, M., Claxton, R.: Network diversity and economic development. *Science* **328**(5981), 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. *Kdd* **96**(34), 226–231 (1996)
- Fekih, M., Bellemans, T., Smoreda, Z., Bonnel, P., Furno, A., Galland, S.: A data-driven approach for origin–destination matrix construction from cellular network signalling data: a case study of Lyon region (France). *Transportation* **66**, 1–32 (2020)
- Frias-Martinez, V., Virseda, J., Rubio, A., Frias-Martinez, E.: Towards large scale technology impact analyses: automatic residential localization from mobile phone-call data. In: 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. *Comput. Environ. Urban Syst.* **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 Soc. Behav. Sci.* **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. *J. Mod. Transp.* **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. *Transp. Res. 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, A., Hamed, M., Sadabadi, K.F.: 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 Highway Administration. <https://www.fhwa.dot.gov/policinformation/hpms.cfm> (2020)
- Horak, R.: *Telecommunications and Data Communications Handbook*. Wiley (2007)
- Houston-Galveston Area Council of Governments. Draft Summary Report: 2008-09 Regional Household Activity/Travel Survey. ETC Institute (2009)
- Hu, P.S., Reuscher, T.R.: Summary of Travel Trends: 2001 National Household Travel Survey (2004)
- Huang, H., Cheng, Y., Weibel, R.: Transport mode detection based on mobile phone network data: a systematic review. *Transp. Res. C Emerg. Technol.* **101**, 297–312 (2019)
- In-The-Moment Travel Study. Revised Report. RSG (2015)
- INRIX Traffic: <http://www.inrix.com/> (2020)

- Jenks, G.F.: The data model concept in statistical mapping. *Int. Yearb. Cartogr.* **7**, 186–190 (1967)
- Kang, C., Liu, Y., Ma, X., Wu, L.: Towards estimating urban population distributions from mobile call data. *J. Urban Technol.* **19**(4), 3–21 (2012a)
- Kang, C., Ma, X., Tong, D., Liu, Y.: Intra-urban human mobility patterns: an urban morphology perspective. *Phys. A* **391**(4), 1702–1717 (2012b)
- Kansas City Regional Travel Survey: Final Report. NuStats, Austin (2004)
- Kearns, M., Valiant, L.G.: Learning Boolean formulae or finite automata is as hard as factoring. Technical Report TR-14–88, Harvard University Aiken Computation Laboratory (1988)
- Kearns, M., Valiant, L.G.: Cryptographic limitations on learning Boolean formulae and finite automata. *J. Assoc. Comput. Mach.* **41**(1), 67–95 (1994)
- Landmark, A.D., Arnesen, P., Södersten, C.J., Hjelkrem, O.A.: Mobile phone data in transportation research: methods for benchmarking against other data sources. *Transportation* **66**, 1–23 (2021)
- Lapham, S.J.: 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, DC (2007)
- Mid-Region Council of Governments 2013 Household Travel Survey. Final Report. Westat, Rockville (2014)
- National Capital Region Transportation Planning Board, Metropolitan Washington Council of Governments. 2007/2008 TPB Household Travel Survey Technical Documentation (2010)
- Nguyen, M.H., Armoogum, J.: Hierarchical process of travel mode imputation from GPS data in a motorcycle-dependent area. *Travel. Behav. Soc.* **21**, 109–120 (2020)
- Nitsche, P., Widhalm, P., Breuss, S., Brändle, N., Maurer, P.: Supporting large-scale travel surveys with smartphones—a practical approach. *Transp. Res. C Emerg. Technol.* **43**, 212–221 (2014)
- Ojah, M., 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., Zhang, L.: Quantifying human mobility behaviour changes during the COVID-19 outbreak in the United States. *Sci. Rep.* **10**(1), 1–9 (2020)
- Pappalardo, L., Simini, F., Rinzivillo, S., Pedreschi, D., Giannotti, F., Barabási, A.-L.: Returners and explorers dichotomy in human mobility. *Nat. Commun.* **6**, 8166 (2015)
- Patterson, Z., Fitzsimmons, K.: Datamobile: Smartphone travel survey experiment. *Transp. Res. Rec. J. Transp. Res. Board* **2594**(1), 35–43 (2016)
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J.: Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **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 (2015)
- Quinlan, J.R.: Induction of decision trees. *Mach. Learn.* **1**(1), 81–106 (1986)
- Safi, H., Assemi, B., Mesbah, M., Ferreira, L.: Trip detection with smartphone-assisted collection of travel data. *Transp. Res. Rec.* **2594**, 18–26 (2016)
- Schmidhuber, J.: Deep learning in neural networks: an overview. *Neural Netw.* **61**, 85–117 (2015)
- Schönfelder, S., et al.: Exploring the potentials of automatically collected GPS data for travel behaviour analysis: a Swedish data source. *Arbeitsberichte Verkehrs-und Raumplanung* **124**, 66 (2002)
- Schrank, D., Eisele, B., Lomax, T.: 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. *Transp. Res. Rec. J. Transp. Res. 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., Koren, T., Wang, P., Barabási, A.-L.: Modelling the scaling properties of human mobility. *Nat. Phys.* **6**(10), 818 (2010a)
- Song, C., Qu, Z., Blumm, N., Barabási, A.-L.: Limits of predictability in human mobility. *Science* **327**(5968), 1018–1102 (2010b)
- Soto, V., Frias-Martinez, V., Virseda, J., Frias-Martinez, E.: Prediction of socioeconomic levels using cell phone records. In: International Conference on User Modeling, Adaptation, and Personalization. Springer (2010)

- Stenneth, L., et al.: Transportation mode detection using mobile phones and GIS information. In: Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information System (2011)
- Stopher, P.R., Jiang, Q., FitzGerald, C.: Processing GPS data from travel surveys. In: 2nd International Colloquium 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. *Transp. Res. C Emerg. Technol.* **16**(3), 350–369 (2008)
- Suykens, J.A., Vandewalle, J.: Least squares support vector machine classifiers. *Neural Process. Lett.* **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. *Transp. Res. Rec. J. Transp. Res. Board* **1972**(1), 38–45 (2006)
- U.S. Department of Transportation, Bureau of Transportation Statistics, *Transportation Statistics Annual Report 2020*. Washington, DC. <https://doi.org/10.21949/1520449> (2020)
- 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)
- Vaughan, J., Imani, A.F., Yusuf, B., Miller, E.J.: Modelling cellphone trace travel mode with neural networks using transit smartcard and home interview survey data. *Eur. J. Trans. Infrastruct. Res.* **20**(4), 269–285 (2020)
- Wang, L. (Ed.): *Support Vector Machines: Theory and Applications* (Vol. 177). Springer (2005)
- Wang, F., Chen, C.: On data processing required to derive mobility patterns from passively-generated mobile phone data. *Trans. Res. C Emerg. Technol.* **87**, 58–74 (2018)
- Wang, B., Gao, L., Juan, Z.: Travel mode detection using GPS data and socioeconomic attributes based on a random forest classifier. *IEEE Trans. Intell. Transp. Syst.* **19**(5), 1547–1558 (2017)
- 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. *Transp. Res. C Emerg. Technol.* **105**, 183–202 (2019)
- Wichita Falls Urban Transportation Study. Summary Report: 2010–11 Regional Household Activity/Travel Survey. ETC Institute (2011c)
- 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., Lee, M.: Synthesis of and statistics for recent GPS-enhanced travel surveys. In: Proceedings of the International Conference on Survey Methods in Transport: Harmonization and Data Comparability, International Steering Committee for Travel Survey Conferences. Annecy, France (2008)
- Wolf, J., Guensler, R., Bachman, W.: Elimination of the travel diary: Experiment to derive trip purpose from global positioning system travel data. *Transp. Res. Rec. J. Transp. Res. Board* **1768**(1), 125–134 (2001)
- Xiao, G., Juan, Z., Zhang, C.: Travel mode detection based on GPS track data and Bayesian networks. *Comput. Environ. Urban Syst.* **54**, 14–22 (2015)
- Xiong, C., Shahabi, M., Zhao, J., Yin, Y., Zhou, X., Zhang, L.: An integrated and personalized traveler information and incentive scheme for energy efficient mobility systems. *Transp. Res. C Emerg. Technol.* **6**, 66 (2019)
- Xiong, C., Hu, S., Yang, M., Luo, W., Zhang, L.: Mobile device data reveal the dynamics in a positive relationship between human mobility and COVID-19 infections. *Proc. Natl. Acad. Sci.* **117**(44), 27087–27089 (2020a)
- Xiong, C., Hu, S., Yang, M., Younes, H., Luo, W., Ghader, S., 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. *J. R. Soc. Interface* **17**(173), 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, DC (2019)
- Ye, Y., Zheng, Y., Chen, Y., Feng, J., Xie, X.: Mining individual life pattern based on location history. In: 2009 Tenth International Conference on Mobile Data Management: Systems, Services and Middleware, pp. 1–10 (2009)
- Zhang, L., Viswanathan, K.: The on-line travel survey manual: a dynamic document for transportation professionals. *Transp. Res. Board* **17**, 66 (2013)

- Zhang, L., Sepehr G., Michael L.P., Chenfeng X., Aref D., Mofeng Y., Qianqian S., AliAkbar K., 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 Trans. Inf. Syst.* **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 Trans. Intell. Transp. Syst.* **18**(8), 2096–2110 (2016)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Mofeng Yang is a Ph.D. Candidate at Maryland Transportation Institute (MTI), University of Maryland, College Park, USA. He received a M.S. in Civil Engineering from University of Maryland, College Park and a B.E. in Traffic Engineering from Southeast University, China. His research focuses on human mobility, large scale geo-spatial computation, big data analytics and traffic simulations and operations. He has published over 15 peer-reviewed journal and conference papers on these topics.

Yixuan Pan is a doctoral student at the University of Maryland, College Park. She holds a Master's degree in Civil Engineering from the University of Maryland, College Park and a Bachelor's degree in Transportation Engineering from the Tongji University, in China. Her main research interests are travel demand analysis and data analytics on mobile device location data and their applications to multimodal transportation engineering.

Aref Darzi is a doctoral student at University of Maryland in United States. He holds a Master's degree in transportation planning and a Bachelor's degree in Civil Engineering both from Sharif University of Technology. His main research interests are big data and transportation, human mobility pattern, travel demand modeling, and travel behavior.

Sepehr Ghader is a transportation and mobility expert, working as an analyst at Citadel. He previously served as the Assistant Director of the Maryland Transportation Institute at the University of Maryland. He received a B.S. in Civil Engineering from Sharif University of Technology, an M.S. degree in Civil Engineering, an M.A. degree in Mathematical Statistics, and a Ph.D. degree in Transportation Engineering from the University of Maryland.

Chenfeng Xiong is an Associate Research Professor in the Department of Civil and Environmental Engineering at the University of Maryland, College Park, and an Associate Professor in the Shock Trauma and Anesthesiology Research (STAR) Center in the School of Medicine at the University of Maryland, Baltimore. He holds an MA degree in Economics and a Ph.D. degree in Transportation Engineering from the University of Maryland. Dr. Xiong's research is focused on travel behavior modeling, transportation systems analysis, modeling and simulation, and its applications in smart mobility and health. He has published over 70 peer-reviewed journal and conference papers on these topics. Dr. Xiong is a recipient of 2019 Outstanding Young Researcher Award of International Association of Public Transport (UITP), Young Initiatives Award of World Conference on Transportation Research Society (WCTRS), and several best paper awards from international conferences.

Lei Zhang is the Herbert Rabin Distinguished Professor of Civil Engineering and Director of the Maryland Transportation Institute at the University of Maryland (UMD). He has published many peer-reviewed journal and conference papers on topics including large-scale transportation systems modeling and simulation, transportation data analysis including survey and passively collected data, travel behavior, transportation economics and policy, and innovative mobility and sustainability solutions.

Authors and Affiliations

Mofeng Yang¹ · Yixuan Pan¹ · Aref Darzi¹ · Sepehr Ghader¹ ·
Chenfeng Xiong^{1,2} · Lei Zhang¹

Mofeng Yang
mofeng@umd.edu

Yixuan Pan
ypan1003@umd.edu

Aref Darzi
adarzi@umd.edu

Sepehr Ghader
Sepehr.gdr@gmail.com

Chenfeng Xiong
cxiong@umd.edu

¹ Department of Civil and Environmental Engineering, Maryland Transportation Institute (MTI), University of Maryland, 1173 Glenn Martin Hall, College Park, MD 20742, USA

² Shock Trauma and Anesthesiology Research (STAR) Center, School of Medicine, University of Maryland, 685 W Baltimore Street, Suite 600, Baltimore, MD 21201, USA