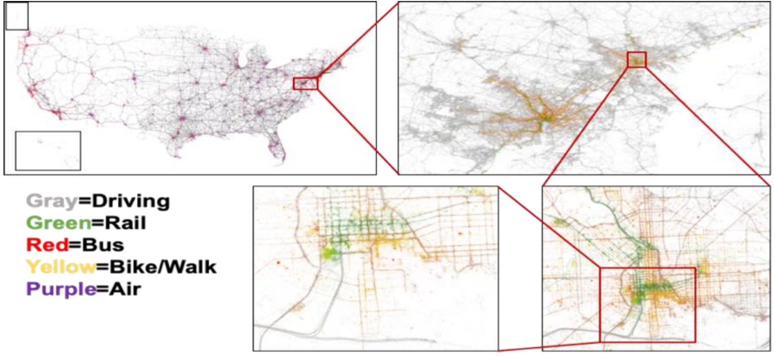
Project Number: 693JJ31750013

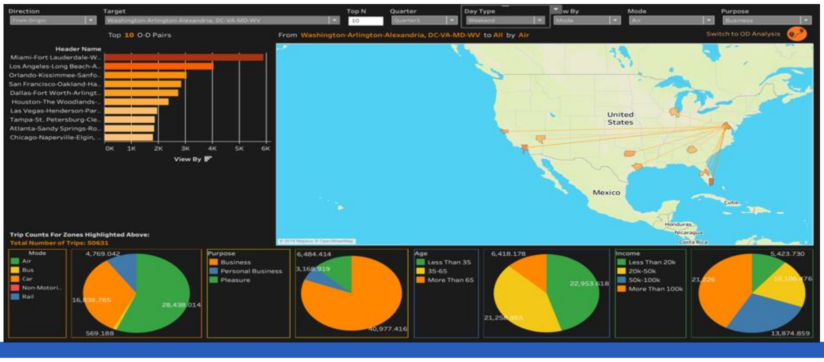
An FHWA Exploratory Advanced Research Program Project

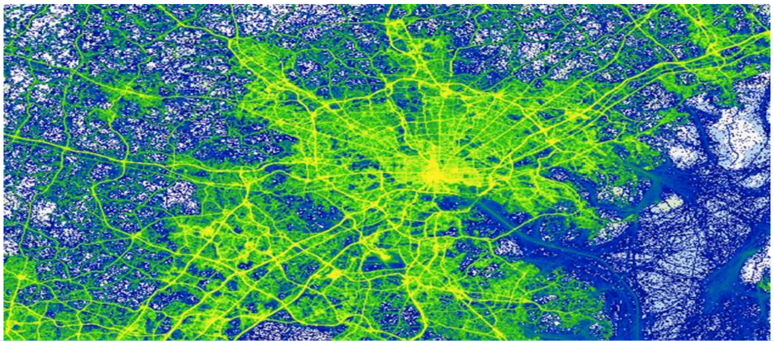
Data Analytics and Modeling Methods for Tracking and Predicting Origin-Destination Travel Trends based on Mobile Device Data

Final Report

June 2020







FHWA COTR:
Patrick Zhang

UMD Principal Investigator:
Lei Zhang

UMD Project Manager:
Sepehr Ghader

UMD Project team:
Aref Darzi, Kathleen Stewart, Junchuan Fan, Yixuan Pan, Mofeng Yang, Qianqian Sun, Aliakbar Kabiri, Guangchen Zhao

Prime Contractor:
University of Maryland (UMD)

Subcontractors:
AirSage, INRIX

Agency Partners:
Maryland Department of Transportation State Highway Administration
Baltimore Metropolitan Council

EXECUTIVE SUMMARY

Background and Project Objective

As part of a Federal Highway Administration's Exploratory Advanced Research (EAR) Program project, the Maryland Transportation Institute (MTI) at the University of Maryland (UMD) worked with data providers and agency partners to explore the feasibility of estimating high-quality passenger and truck travel origin-destination (OD) tables from mobile device data at the national and statewide/metropolitan levels. As mobile device data are becoming increasingly available, the UMD-led research team focused on both the evaluation of raw mobile device data and the development of OD products. The team was dedicated to developing transparent methods that can build user confidence in both the raw data and relevant products. Mobile device data can be collected with minimal public burden and cost, and can supplement traditional methods of data collection to offer a more comprehensive understanding of changes in travel trends. Mobile device datasets provide more timely information, have a larger sample size, cover all modes of transportation, and requires no data collection burden on users. On the other hand, mobile device datasets are known to have a biased sample; miss key travel information such as trip purpose, travel mode, and socio-demographics; and originate from different technologies and sources, which must be addressed before they can be used to produce quality OD tables.

Data and Products

To test the feasibility of using mobile device data to analyze OD patterns, the UMD-led research team collected raw data from a variety of sources, including global positioning systems (GPS) embedded in phones, cars, and trucks; communication data between cellphone towers and mobile devices; and data from location-based services (LBS) embedded in a large number of smartphone apps. The research team also examined the OD patterns of truck movements in addition to passenger travel based on truck GPS data. The research team collaborated with established data providers (i.e., AirSage, INRIX, and StreetLight as a sub to INRIX) to collect this data. In order to protect user privacy, the input data used in the research are anonymized and do not include personal information about individual users. The end products from the project include:

- National MSA-to-MSA-level passenger OD tables by day type, trip purpose, travel mode, and socio-demographics
- National MSA-to-MSA-level truck OD tables by day type, and vehicle weight class
- Sample metropolitan TAZ-TAZ (traffic analysis zone)-level passenger OD tables by day type, time of day, trip purpose, travel mode, and socio-demographics
- Sample metropolitan TAZ-TAZ-level truck OD tables by day type, time of day, and vehicle weight class

Research Approach and Methodology

After gaining access to and checking the quality of raw mobile device location data from several data provider partners, UMD developed a variety of computation algorithms for raw data cleaning and quality enhancement, trip identification, imputation of missing information (e.g., travel modes including air, rail, bus, car, walk, and bike; trip purposes; and socio-demographics), and sample

extrapolation and weighting. The research team first cleaned the raw location data to remove duplicate and unreasonable records and converted it from location points with time stamps to individual trips for subsequent analysis. The team then developed algorithms to impute travel information that wasn't directly collected from the raw mobile device data. For instance, the research team imputed travel modes by studying movement patterns and integration with a multi-modal transportation network; imputed trip purposes by studying daily travel patterns and home/work locations and integration with point-of-interest data; and imputed travelers' socio-demographic information such as income and age by studying travel patterns and integration with census information. Weight algorithms were developed and implemented to expand the sample and address biases. The OD products at the national and metropolitan levels also went through several validation tests based on independent passenger and truck travel data from household travel surveys, Highway Performance Monitoring System, traffic counting stations, the National Transit Database, airline DB1B and T100 datasets, Freight Analysis Framework, and more.

To improve data transparency, the UMD-led team developed a Raw Data Sandbox as an extra step beyond the project scope, which includes anonymized raw mobile device location data from several data providers. Data product users and researchers may use this data sandbox to better understand the raw data properties and even test their own computational algorithms in the data sandbox. For methodology transparency, all computational algorithms developed by UMD for this FHWA EAR project are described in detail in this project report. In addition, source codes will eventually be open-source and shared with the broader research and data user communities.

Conclusions

This project demonstrated that it is feasible to produce quality passenger and truck travel OD tables at the national and statewide/metropolitan levels from mobile device data. Computational algorithms have been developed and implemented to address known issues with mobile device data such as sample bias and missing trip and traveler information. Validation results show that OD tables based on mobile device data are consistent with control totals derived from established independent data sources. Future research may focus on establishing national standards or guidelines on raw data quality, data and method transparency, computational algorithm performance, and validation targets. FHWA and other agencies may consider integrating OD products from mobile device data into their existing data programs and business processes.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Research Background

Billions of location data points are generated each day from mobile devices. These include cell phone record data, global positioning system (GPS) location data from sensors in navigators, phones, and in-vehicle systems, and smartphone app location data from location-based services (LBS) in smartphones (**Figure 1**). In the past decade, mobile device location data have become available through various data providers for transportation planning and operations, as they possess some advantages over traditional data sources for travel monitoring and analysis (e.g., household travel surveys and origin-destination (OD) tables estimated from link-based counts). Notable features of mobile device location data are continuous tracking, large sample size in terms of both devices and trips, and passive data collection. Typically, mobile device location data records contain latitude and longitude coordinates and time stamps. Certain trip and traveler information (e.g., trip ends, travel mode, purpose, and the socio-economic and demographical characteristics of travelers) is not directly available from mobile devices and must be imputed with machine learning, data fusion, and statistical methods.

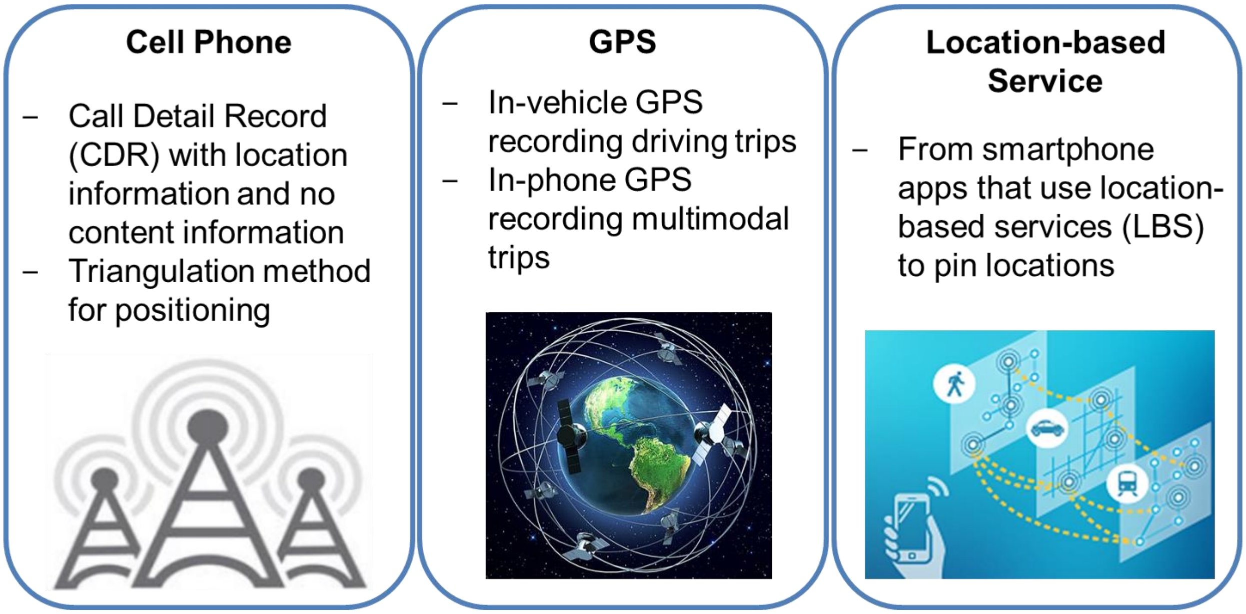
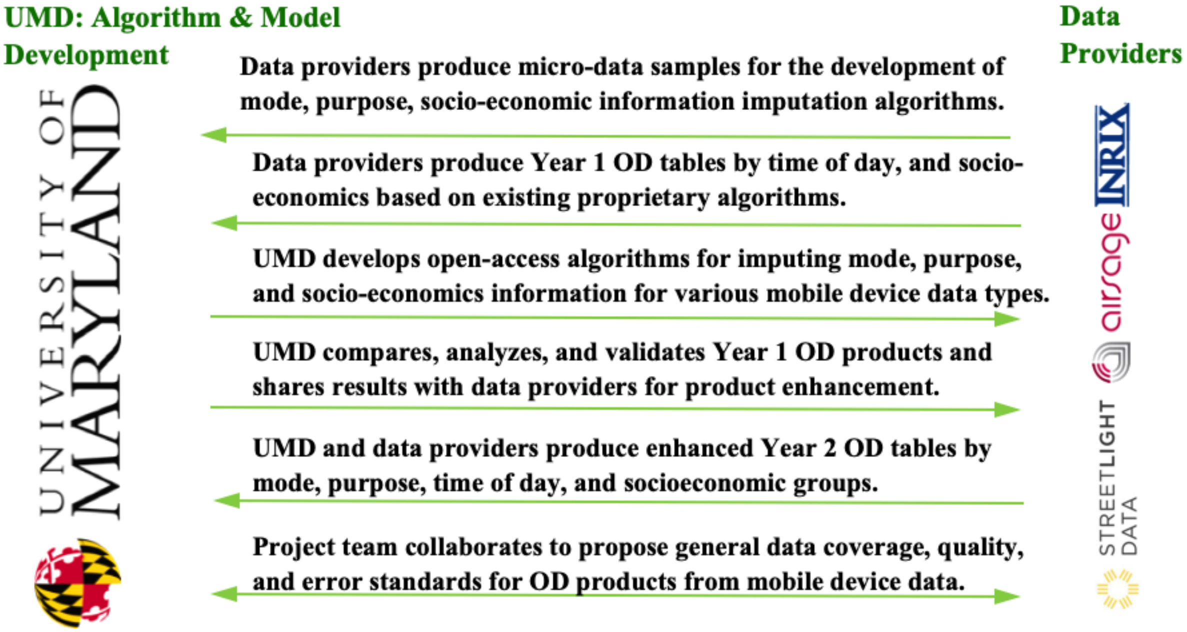

Cell Phone
– Call Detail Record (CDR) with location information and no content information
– Triangulation method for positioning
GPS
– In-vehicle GPS recording driving trips
– In-phone GPS recording multimodal trips
Location-based Service
– From smartphone apps that use location-based services (LBS) to pin locations

Figure 1. Sources of mobile device location data

On the application side, mobile devices have been primarily employed as probes for the estimation of aggregate travel and traffic patterns, such as OD tables (Calabrese, et al., 2011), mode share (Doyle, et al., 2011), travel time (Liu and Ma, 2008), speed (Bar-Gera, 2007), and traffic volume (Caceres, et al., 2012). In some cases, state, metropolitan, and local transportation agencies have used mobile device location data because they fill an important travel data gap: a lack of recent OD information, especially OD information of external trips. Several state departments of transportation (e.g., District of Columbia, Maryland, Ohio, and Virginia) have acquired OD tables at the census tract level or higher from mobile device data providers. Researchers and practitioners have developed algorithms to identify trip ends, time of day, purposes (usually home, work, and other purposes), and socio-economic and demographical characteristics of travelers. Statistical expansion methods based on device penetration rates are often used by data providers to extrapolate the mobile device sample to the entire population. Validation of travel information derived from mobile device data has been limited to a handful of studies that compare mobile device data products with traditional travel surveys and regional control totals.

The main motivation of this research project was to address four main limitations of OD products. The first limitation was that no person and truck travel OD data at the national level based on mobile device location data existed. Second, there is a lack of timely data on OD travel patterns at the MPO levels. The third limitation was that the existing OD products from passively collected mobile device location data were not multi-modal and the fourth was the lack of transparency of raw data and computational algorithms in commercially offered OD products at the statewide and MPO levels based on mobile device location data. The technology developed in this EAR project has sufficiently addressed these four problems by producing multi-modal national person and truck travel OD products, producing sample multi-modal MPO OD products, providing a raw data sandbox for improved data transparency as an extra step beyond the project scope, and open-sourcing all computational algorithms developed by the research team for method transparency. This project helped advance the state-of-the-art and practice over the project years. UMD has worked closely with three commercial data provider companies in this project, AirSage, INRIX, and StreetLight Data (as a sub to INRIX). Through an iterative process, the project team has improved the technology so it can be integrated into the companies' business processes for producing OD products to their clients. One commercial data provider company is already using algorithms developed by this project in their own OD production for other clients. **Figure 2** summarizes the collaboration plan between the UMD team and the data provider partners.



Data providers produce micro-data samples for the development of mode, purpose, socio-economic information imputation algorithms.

Data providers produce Year 1 OD tables by time of day, and socio-economics based on existing proprietary algorithms.

UMD develops open-access algorithms for imputing mode, purpose, and socio-economics information for various mobile device data types.

UMD compares, analyzes, and validates Year 1 OD products and shares results with data providers for product enhancement.

UMD and data providers produce enhanced Year 2 OD tables by mode, purpose, time of day, and socioeconomic groups.

Project team collaborates to propose general data coverage, quality, and error standards for OD products from mobile device data.

Figure 2. UMD collaboration with the three main data providers on the team

1.2. Research Objectives

The main project goal was to explore the potential of using mobile device data to produce quality person and truck travel origin-destination (OD) data with transparent approaches to increase and improve data user confidence. The detailed project objectives were as follows:

- **Explore the potential of producing national person travel origin-destination (OD) tables from passively-collected mobile device data (e.g., cell phone, GPS, LBS):** The team achieved this objective and showed the feasibility of producing national OD tables from mobile device data. The team also evaluated the national products through a comprehensive validation effort and showed that mobile device data can be utilized to produce high-quality OD products.

- **Explore the potential of segregating person travel OD data by mode, purpose, time-period, socio-economic, and demographical variables:** The team achieved this objective by producing person OD tables that include all the segregating attributes. The team not only validated the algorithms involved in producing the segregating attributes separately, they also validated these attributes in the end products and showed satisfactory results. All algorithms involved are open-source and can be used by the public to increase transparency.

- **Explore the potential of generating truck travel OD tables that are segregated by time-period and vehicle weight class:** The team achieved this objective by producing truck OD tables that include all the segregating attributes.

- **Explore partnership with universities and data providers to deliver needed data and information:** The team achieved this objective by demonstrating a successful collaboration with the project partners. The team successfully collected required raw data and state-of-the-practice OD products from all partners and developed data quality methods and statistical/machine learning algorithms applicable to their data. This collaboration not only benefited the partnering data providers, but also led to an improvement in the state-of-the-practice followed by other data providers.

1.3. Research Contributions

The contributions of this research can be summarized as follows:

- **Producing the first national OD tables from mobile device data for person and truck travel:** While mobile device data were used to produce OD tables at local and MPO levels, no national OD tables from mobile device data existed before this project. With major research and development efforts, this project produced the first set of national OD tables from mobile device data for person and truck travel. This effort contributed to the state-of-the-practice, where national products are now available from commercial data providers.

- **Producing the first multimodal OD tables from mobile device location data at both national and MPO levels:** Before this project, OD tables were limited to all trips or driving trips only. The literature included state-of-the-art research efforts to impute travel mode from mobile device data, but none of these efforts were applied in a large scale to produce OD tables. The research team successfully developed, applied, and validated various mode imputation algorithms, which led to the first multi-modal OD tables from mobile device data for person travel.

- **Elevate the current methodologies involved in producing OD tables from mobile device data, such as trip identification, mode imputation, and weighting:** The research team developed and tested various statistical and machine learning algorithms to produce the OD tables. In addition to validating the algorithms at both aggregate and individual levels, the end products were also validated and showed satisfactory results.

- **Improving the transparency of the entire process by open-source methodologies, comprehensive validation, and raw data sandbox:** One major limitation of mobile device products has always been the lack of transparency. For legitimate business reasons, commercial data providers rarely share detailed information about their raw data, their methodologies, or validation results. As a non-profit university research team, our team made all possible efforts to promote transparency. Researchers can read open-source computational algorithm details published by the research team to better understand how the OD tables are produced from raw data and to improve their own algorithms. They can also use the raw data sandbox to better understand raw data and to test their own computational algorithms.

1.4. Research Product Summary

Table 1 presents a summary of the project's research products.

Table 1. Product description summary

Product	Description	Zone Structure
National-Level Person OD	Weighted 2017 long-distance person OD for interzonal trips by day type, socio-demo., purpose, and mode	MSA Level
National-Level Truck OD	Weighted 2017 long-distance truck OD by day-type and vehicle weight class	MSA Level
MPO-Level Person OD	Weighted 2017 all-trips person OD by day type, socio-demo., purpose, mode, and time-of-day for the Baltimore metropolitan region	TAZ level
MPO-Level Truck OD	Weighted 2017 truck OD by day type, vehicle weight class, and time-of-day for the Baltimore metropolitan region	TAZ level

As seen in the table, the national OD products are at the metropolitan statistical (MSA) level, and only include long-distance trips. For each state, the geography includes all MSAs located in that state, with the remaining parts of the state that do not belong to any MSA forming one combined non-MSA zone. In total, the project's national zone structure includes 441 zones covering the continental U.S. plus Hawaii and Alaska. National person OD tables include trips between all zones while national truck OD tables only cover the continental U.S. Furthermore, national person OD tables only include interzonal trips, while truck OD tables also include intrazonal trips. These differences between the national person and truck OD tables are due to innate differences between the original data sources. Truck OD tables are produced from in-vehicle GPS devices, while person OD tables are produced from LBS data.

In this project, long-distance trips were defined as trips longer than 50 miles. This definition was selected based on its popularity and simplicity. The data processing methods implemented by the research team can be applied to any definition of long-distance trips, as all trips are first processed and long-distance trips are later filtered. To showcase the capability of producing OD tables for a more disaggregated zone structure, the research team selected the Baltimore metropolitan area as the case study for producing regional all-trip OD products. The zone structure for the MPO OD tables was directly provided by the Baltimore Metropolitan Council (BMC). The zone structure includes 2922 TAZs. **Figure 3** shows the zone structure for the national and the MPO products.

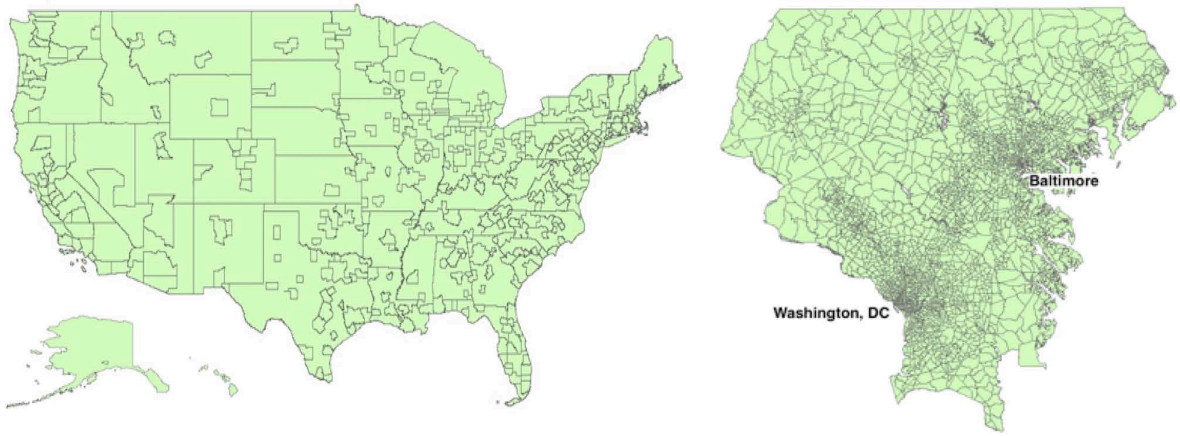


Figure 3. Zone structure for the national and MPO OD products

National OD tables are separated into three categories for day type: average day, average weekday for Monday to Friday, and average weekend for Saturday and Sunday. Holidays are not excluded in any of the OD products. National person OD products include three purposes: business, personal business, and pleasure. Purpose categories are based on the 1995 American Travel Survey purposes, aggregated to three general categories (business, personal business, and pleasure). Personal business includes school-related activities, personal, family, or medical purposes. Business category includes business, combined business/pleasure, convention, conference, and seminar purposes. National person OD products also include a mode attribute covering the following modes: driving, bus, rail, and air. MPO OD tables are separated into three categories for day type: average day, average weekday for Monday to Friday, and average weekend for Saturday and Sunday. Holidays are not excluded in any of the OD products. MPO OD products include the following five time-of-day intervals: all day, morning peak, 6am to 10am; mid-day, 10am to 3pm; afternoon peak, 3pm to 7pm; and night, 7pm to 6am. MPO person OD products include trip purpose based on the following categories: home-based work, home-based other, and non-home-based. MPO person OD products also include a mode attribute covering the following modes: driving, bus, rail, air, and non-motorized.

Both national and MPO person OD tables are also divided by socio-demographic information based on the following attributes:

- Income: less than $20,000, $20,000 to $50,000, $50,000 to $100,000, more than $100,000
- Age: young, less than 35; middle-age, 35 to 65; senior, older than 65
- Gender: male, female.

2. LITERATURE REVIEW

This section reviews the related mobile device data applications in transportation, computer science, and geography literature. The first sub-section provides a summary of different types of mobile data and their applications, and the second part provides a summary of related computation/statistical methods that are applied to mobile device data.

2.1. Mobile Device Data and Their Applications

Travel surveys have been a key part of transportation planning and have been widely used around the world since the 1950s. The first approach for conducting surveys was the face-to-face approach; but due to technology advancement and the high cost and safety issues associated with face-to-face surveys, they have been mainly replaced with mail-out/mail-back, telephone, or web surveys. Later, travel surveys took advantage of technology developments and the computer started playing a role in collecting surveys. Computer-assisted telephone-interview (CATI), and computer-assisted self-interview (CASI) approaches have been used to provide respondents a better understanding of the questions and for completing the questionnaires more effectively (Shen and Stopher, 2014; Wolf, et al., 2001). However, some issues, such as an underreported number of generated trips and inaccurate trip times, still remain as inherent disadvantages of these kinds of travel surveys (McGowen and McNally, 2007). These limitations are among the key reasons why the use of mobile device data to collect travel information is gaining popularity.

The most widely-used type of mobile device data is the data coming from GPS technology, which is capable of recording accurate information including location, time, speed, and possibly a measure of data quality (Stopher, et al., 2008). Because of its ability to supplement traditional travel surveys, researchers in the mid-1990s began investigating the possibility of using GPS data to improve survey results and also test GPS data accuracy (Murakami and Wagner, 1999; Sermons and Koppelman, 1996; Wagner, 1997). The earliest version of GPS devices was only capable of recording vehicle movements since the device was electrified by the vehicle's battery. As the technology improved in the early 2000s, GPS devices became smaller and lighter and could be used with detached batteries (Stopher, et al., 2008). These improvements solved the problem of only capturing the vehicle's movements and led to wearable GPS devices (2nd generation, 3rd generation). Although the new generation of GPS devices provided a broad set of applications, these devices still have disadvantages, such as signal loss and the fact that respondents might forget to take the device on their trips (Gong, et al., 2014).

GPS technology can improve the temporal and spatial accuracy of travel surveys significantly by recording the exact origin and destination location as well as the exact trip start and end times. The first attempts of utilizing GPS technology in travel diaries can be dated back to a 1996 Lexington study sponsored by FHWA (Wagner, 1997), the 1997 Austin (TX) travel survey (Pearson, 2001), the Netherland's travel survey (Draijer, et al., 2000), a 1998 truck study in California (Wagner, et al., 1998), and the 1999 travel survey in Quebec City (Doherty, et al., 2001). Furthermore, a handful of experiments also showed the feasibility of collecting travel data via GPS devices, either a handheld electronic travel diary (ETD) with a GPS or a passive in-vehicle GPS system, to complement traditional household travel surveys (Doherty, et al., 2001; Draijer, et al., 2000; Pearson, 2001; Wolf, et al., 2000; Wolf, 2000; Yalamanchili, et al., 1999). These early studies emphasized the advantages of a GPS survey in collecting misreported or underreported trips from traditional surveys and documenting more detailed travel activities. Meanwhile, they raised several concerns. For instance, in the case of ETD with GPS, users might not carry the device when they consider it a burden; or in the case of a passive in-vehicle GPS system, the device is only able to capture driving trips and lacks a user interface to validate the trip information. By the beginning of the 21st century, U.S. President Clinton discontinued the selective availability (SA) feature of the publicly available navigation, which paved the road for the use of GPS technology. SA was an intentional degradation of public GPS devices' accuracy up to 100 meters for real-time non-military users and was by far the biggest cause of errors in GPS positioning (Zito, et al., 1995).

More research has been done since the practicality of GPS travel surveys was demonstrated. Many countries used GPS for several purposes, such as surveying and traffic safety evaluation, including Japan, Canada, the Netherlands, France, South Africa, Switzerland, Australia, the U.S., Austria, China, and the UK. Itsubo and Hato (2006) used GPS devices to compare GPS records and travel diaries in Matsuyama, Japan. They surveyed 31 respondents in a 5-day period using GPS-equipped mobile phones with a 30-second frequency. Papinski, et al. (2009) utilized a GPS survey for route choice studies. In their studies, they used smartphones equipped with a GPS receiver to survey 31 people for 2 days along with a pre-interview and a web-based revealed preference survey. Marchal, et al. (2008) used dedicated GPS devices in France and surveyed a 9% sub-sample of their national travel survey for 7 days to complement the national travel survey. In South Africa, dedicated GPS devices have been used to survey 100 respondents during 14 days in order to assess the reliability of GPS surveys (Krygsman and Nel, 2009). The GPS device frequency was set to record data every second and the survey was complemented by a two-day travel diary. Dedicated GPS devices were used in Switzerland to find out whether participants passed certain billboards (Schuessler and Axhausen, 2009); 4882 people participated in this survey and the average period of the survey was about 6.6 days. In Australia, dedicated GPS devices were used to record data every second from 130 households to monitor their travel behavior changes (Stopher, et al., 2013). The collection period of the data was about 15 days for 6 waves starting from 2007 to 2012. In Ohio, dedicated GPS devices were used to conduct a GPS-only household travel survey (Stopher and Wargelin, 2010). The survey used random sampling and a GPS-only survey was used with a web-based revealed preference survey; 2059 households were surveyed for 3 days and the frequency of the data recording was set to one second. In Austria, dedicated GPS devices were used to test and evaluate the integration of new technologies for a mobility survey purpose (Kohla and Meschik, 2013). A survey was conducted using random sampling for 4 different groups: passive GPS-only, active GPS-only, GPS with diary, and diary only. A total number of 235 people were surveyed for 3 days. In 2013, as a part of research on travel chaining and sustainable mobility, 54 households were surveyed using GPS devices with a data frequency of 1 second for 3 to 5 days. The survey was conducted by random sampling from the Danish national travel survey, and the GPS survey was complemented with a one-day travel diary (Rasmussen, et al., 2013). Bohte and Maat (2009) and Bohte, et al. (2007) utilized a GPS travel survey to investigate the relationship between residential self-selection and travel behavior.

A GPS-enabled mobile phone is a more convenient and less expensive replacement for a handheld ETD with GPS. The influence of mobile phone location services on intelligent transportation systems is discussed by Zhao (2000). Further studies showed the feasibility of utilizing a GPS-enabled cell phone rather than a dedicated in-vehicle GPS system to monitor locations and movements (Amin, et al., 2008; Byon, et al., 2007; Guido, et al., 2012; Wiehe, et al., 2008; Work, et al., 2009; Zhou, et al., 2005). Smartphones have become a useful tool for conducting surveys (Bierlaire, et al., 2013; Xiao, et al., 2012). Cottrill, et al. (2013) shared their experience in designing a smartphone-based mobility survey, which can provide a better user interface in comparison with GPS-based travel surveys. Since GPS can offer precise locations, the access to individual-level mobile GPS trajectories is more restricted. There are several private sector companies that generate aggregate level location data to reveal travel demand, such as INRIX, StreetLight Data, AirSage, etc. (Leber, 2013). Hard, et al. (2017) compared the cellphone-, GPS-, and Bluetooth-derived OD data in order to determine whether cellphone and GPS data can be used in TxDOT's external survey. The study focused on external-external trips and used one month of cellular data from AirSage and three months of pre-processed GPS data from INRIX. A TTI-collected Bluetooth dataset was considered as the benchmark and their results showed an under-estimation of external-external trips in cellphone data and commercial vehicle bias in GPS data.

Since mobile phones—and later, smartphones—have gained in popularity, investigations into the individual-level mobility patterns have become more practical. The great value of various emerging data sources has been revealed too, including call detail record (CDR), cell phone GPS data, social media location-based services, etc. CDR provides details on calls and messages, such as timestamp, duration, and location(s) of routing cell tower(s) (Horak, 2007). 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 (Pappalardo, et al., 2015; Song, et al., 2010; Song, et al., 2010; Çolak, et al., 2016). CDR is also applied to other research topics such as social network, residential location, socioeconomic level, etc. (Eagle, et al., 2010; Frias-Martinez, et al., 2010; Soto, et al., 2011). Despite the large volume of data, CDR is limited by its spatial resolution, which is determined by the density of cell towers. However, on the positive side, CDR data require less advanced phones and should raise less concern about the user privacy.

Another source of mobile device data is social media location data, in which spatial information can be implied in the posted text or the uploaded picture other than being directly recorded. This data can help enhance the contents of geographic and spatial data. Flanagin and Metzger (2008) included the photo-sharing site, Flickr, in their discussion about volunteered geographic information (VGI). De Choudhury, et al. (2010) tried to automatically generate travel itineraries for popular touristic cities based on the photo streams uploaded to Flickr. They explored where and when travelers were by mining a large number of shared photos with timestamps. Sui and Goodchild (2001) developed the concept of "(social) media as GIS." They illustrated that location-based social networking sites can act as a GIS, since they provide users' locations with timestamps. Naaman (2011) studied the four aspects of geographic information that can be derived from social awareness streams (SAS) data, including districts, landmarks and attractions, paths (and itineraries), and activities. Twitter, Facebook, the photo-sharing site Flickr, and the Foursquare location and presence-sharing service are all counted as SAS platforms. Riederer, et al. (2015) collected a two-year public photo metadata from Instagram. They revealed the potential of social media location data in two ways: first, they demonstrated that the human mobility patterns drawn from photo-sharing networks are comparable with those from CDRs; they also showed that an individual’s ethnicity could be predicted solely based on the location data. Furthermore, there are studies utilizing such data to inspire new location-based services (Bao, et al., 2016), predict the next location to visit (Liu, et al., 2016), link users across domains (Riederer, et al., 2016), identify user’s home location (Gu, et al., 2016), propose possible activity companion (Liao, et al., 2016), etc.

The use of passively collected location data is not limited to the transportation field. Troped, et al. (2008) employed GPS and accelerometer data to predict the physical activity mode, such as walking, running, biking, or driving an automobile. Gilbert and Karahalios (2009) utilized social media data to measure and predict the tie strength between social media friends. Soto, et al. (2011) tried to predict the socioeconomic levels of a population based on the aggregated CDR data. De Montjoye, et al. (2013) tracked human mobility traces and concluded that they are highly unique, which draws discussion on the privacy protection of individuals.

2.2. Computation and Statistical Methods

2.2.1. Trip Information Imputation

Along with the development of mobile device data, a great number of attempts have been made to derive the travel information from raw data. This information includes the unobserved characteristics of trips such as trip purpose and travel mode, in addition to the characteristics of travelers. Gong, et al. (2014) conducted a literature review on the methodologies for deriving personal trips from GPS data. Four processing procedures are discussed including data error recognition, trip identification, travel mode detection, and trip purpose inference. The potential of utilizing GPS trace data for travel behavior analysis was evaluated by Schönfelder, et al. (2002). They tried to post-process the data to identify the drivers, trip ends, stops, trip purposes, and the potential to construct all-mode activity patterns using driving GPS records. Chung and Shalaby (2005) developed a map-matching algorithm to identify the roadway links traveled with the GPS data collected by GPS traces and a GIS database. Built upon that, a rule-based model was constructed to detect the travel mode configuration including predefined multimodal patterns. An enhanced framework was later proposed by Tsui and Shalaby (2006) that first applied a rule-based model to segment trips by mode transfer point (MTP) and then used a fuzzy logic-based algorithm to identify mode within trip segments. Another research on trip identification and mode detection is by Schuessler and Axhausen (2009), which employed a fuzzy logic approach to detect the mode followed by a reasonableness check. They also highlighted the model’s capability in dealing with a large sample and eliminating manual intervention. Gonzalez, et al. (2010) developed a smartphone app, TRAC-IT, in which a neural network algorithm for mode detection was embedded. The multi-layer perceptron took speed, acceleration, and estimated horizontal accuracy among the input variables. Zhang, et al. (2011) proposed a multi-stage algorithm. The three mode classes (walk, bike, motorized vehicles) are identified in the first stage by speed, acceleration, etc.; in the second stage, the detailed modes under motorized vehicles are identified using a Support Vector Machine (SVM) method. Gong, et al. (2012) constructed a GIS-based algorithm to impute the travel mode from a large GPS data set from New York City, a complex urban environment where the urban canyon effects and the multimodal transportation network need more attention. Nitsche, et al. (2014) mainly utilized the acceleration data collected by smartphones to automatically reconstruct trips. They also employed a Discrete Hidden Markov Model (DHMM) to compute the travel modes.

Compared to mode detection studies, fewer papers have tried to address the trip purpose imputation. Earlier attempts used land-use information or a combination of land-use and person socio-demographic information to impute the trip purpose (Bohte and Maat, 2009; Stopher, et al., 2008; Wolf, et al., 2001; Wolf, et al., 2004). Wolf, et al. (2004) utilized geographic information system (GIS) land-use data and used a rule-based method to detect the trip purposes and open the discussion about trip-end identification in a GPS processing system. They used 10 categories for trip purposes and based on a CATI-based recall survey, they reported ten wrong purposes out of 151 trips. Their model was not successful in identifying the pickup/drop-off trips mostly due to the problems with telephone interviews. Stopher, et al. (2008) suggested that personal information can be used in purpose imputation in order to improve accuracy. In their study, additional information was collected from the respondents such as their home, workplace, or school locations, as well as the most frequently used grocery stores. Using a rule-based method, they showed that additional information improved the detection accuracy. Other research also tried to detect the trip purpose by implementing a rule-based system based on GPS information, GIS, and personal information (Bohte and Maat, 2009; Chen, et al., 2010; Cottrill, et al., 2013). Bohte and Maat (2009) inferred trip purpose using GIS information other than home and work locations. The travel mode was determined considering both speed and transit routes. Elango and Guensler (2010) tried to identify trip purposes (home, work, maintenance, discretionary, and multipurpose) based on the home/work locations and the closest POI to trip ends. Huang, et al. (2010) developed an algorithm for activity identification incorporating spatial-temporal POIs' attractiveness (STPA). STPA not only addressed the static attractiveness of business but also added a dynamic factor to demonstrate the variation due to the time-of-day, in which the business-related activity usually happens. In addition to the rule-based method, the probabilistic method has been used for trip purpose imputation (Chen, et al., 2010, Wolf, et al., 2004). In these studies, for each trip, the probability for all trip purposes are calculated based on GPS and land-use information and the trip purpose is estimated by the calculated probability.

By advancements in machine learning algorithms, researchers started to impute trip purpose using machine learning. Griffin and Huang (2005) used the decision tree method to detect trip purpose based on trip stop length and time of trip ends. Their work showed that these attributes can only be used for identifying work and school trips. McGowen and McNally (2007) incorporated the classification tree method and the discriminant analysis to detect purposes using detailed GIS information including locations for point of interest. Their results showed no significant difference between the two methods. Deng and Ji (2010) proposed a decision tree algorithm employing land-use information, socioeconomic information of respondents, and a set of spatiotemporal indices of travel to detect trip purpose. Liu, et al. (2013) imputed activity purposes based on mobile phone call locations and a set of machine learning algorithms. Shen and Stopher (2013) further included tour type identification in their study. Kim, et al. (2014) developed a learning model to impute the activity associated with given stops using data collected by a smartphone-based travel survey. Oliveira, et al. (2014) compared nested multinomial logit and decision tree models in terms of performance in imputing trip purpose. They first categorized household members into eight person-types and then incorporated GPS travel data to impute trip purposes. Ermagun, et al. (2017) utilized Google Place data to study real-time trip purpose prediction. They also found that a random forest outperforms a nested logit model. More studies have discussed the performance of machine learning methods in trip purpose imputation (Gong, et al., 2017; Lu and Zhang, 2015; Montini, et al., 2014; Xiao, et al., 2016).

The literature on imputing socio-demographic information based on mobile data is even more limited. Lu and Pas (1999) demonstrated the relationship between socio-demographics, activity participation, and travel behavior through a structural equation model. Although the paper focused on the direct and indirect effects of socio-demographics on travel behavior, such as the number of trips per day, it inspired the possibility of studying the problem in a reversed way, which is to infer travelers' socio-demographics based on their travel behaviors. Altshuler, et al. (2012) tried to include some indirect location features (numbers of different cell tower IDs and different Wi-Fi network names) to impute individual attributes such as their ethnicity, whether they are a student, and whether they are U.S.-native.

Auld, et al. (2015) studied whether demographic characteristics of travelers could be derived from their travel behavior. Their method can be divided into two parts: person type clustering based on the similarity of travel patterns; and demographics modeling under each person-type, including education, age, gender, license, and household type (defined by household size, number of vehicles, and presence of a child). They used various models and algorithms to impute different attributes: partial decision tree classification algorithm (PART) for person type and license possession; nested logit for education; ordinal logit for age categories; binary logit for gender; and decision tree for the household type. Their framework is restricted by several assumptions; for instance, the GPS trace data must cover at least one full day of travel and the home/work/school locations must be available. Such assumptions may not be fulfilled in some prevailing data sources.

Zhong, et al. (2015) did a similar study for a larger number of users and their location check-ins through a social network. The main demographic characteristics considered were gender, age, and education background. As the location check-ins are not continuous, the feature set was composed of POI and temporal information. They also compared several methods for each response variable type, including logistic regression, support vector machine (SVM), neural networks, etc. Riederer, et al. (2015) aimed to infer the demographics (ethnicity and gender) from people's location data collected by Instagram. They utilized a simple Bayesian inference method and compared the model performances with or without auxiliary data (census data and surrounding venue data from Foursquare). Roy and Pebesma (2017) inferred gender first and then age group under each gender type based on anonymized mobile phone GPS trajectories. For gender imputation, they chose a supervised learning approach of linear discriminant analysis (LDA) and for the age groups, they chose a decision-tree based classification approach. They constructed the feature set with trip-based information and POI data as well as the frequently visited places they discovered. Kosinski, et al. (2013) used Facebook "likes" to predict several sensitive personal attributes, such as sexual orientation, ethnicity, religious and political views, etc.

2.2.2. Imputation Methods

The missing information in the mobile data is usually treated as nominal variables. Therefore, the imputation of such responses is modeled as a nonlinear classification problem. Feng and Timmermans (2016) summarized and compared the algorithms applied to mode detection, including naive Bayesian, Bayesian network, logistic regression, multilayer perceptron, support vector machine, decision table, and decision tree. Those methods are also dominant in the imputation of trip or activity purposes.

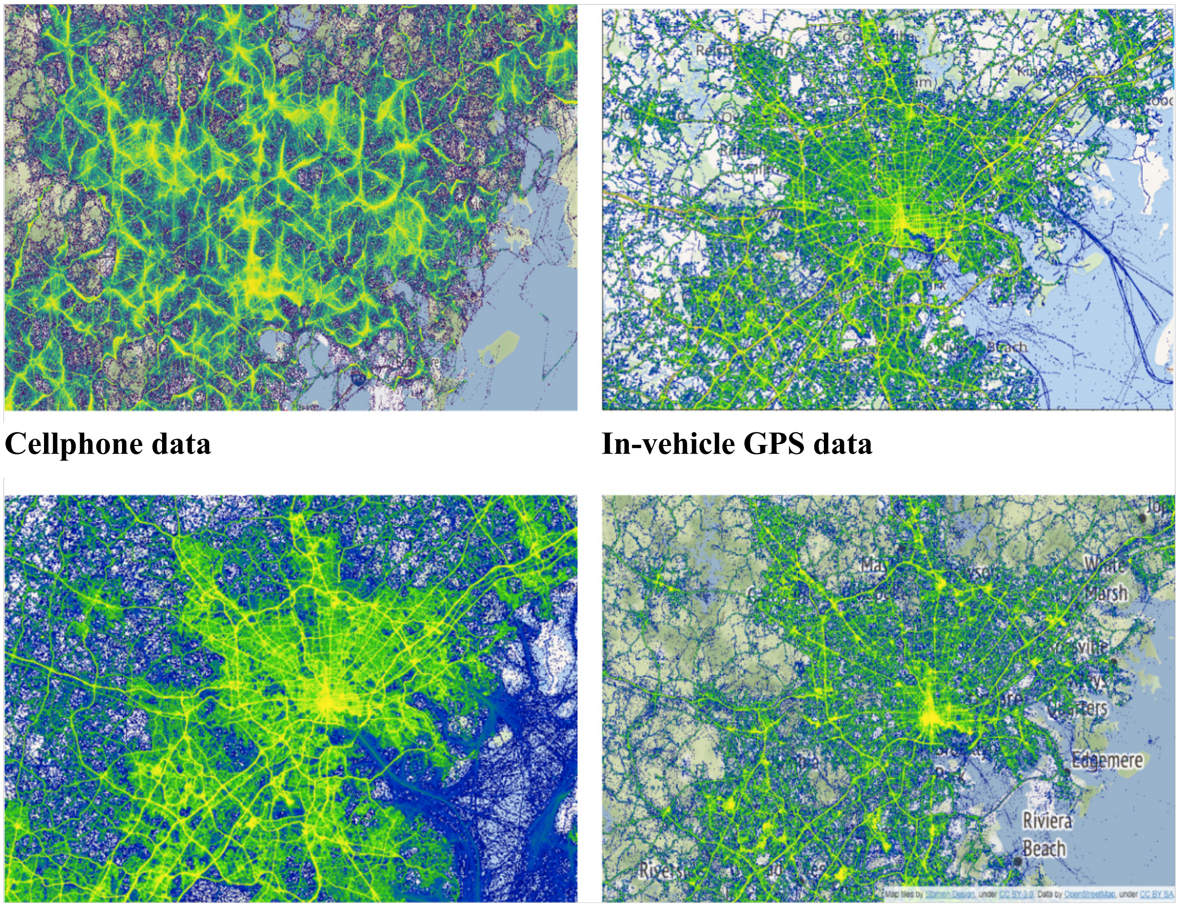
The naive Bayesian method is mainly based on the Bayes' rule, where all predictors are assumed to be independent of each other. The Bayesian network relaxes the assumption and considers the joint probability of an attribute with its parent attributes, but the joint probability distribution could hardly be employed when the dimensions of predictors and their possible values exceed two. To simplify the risk model, the conditionally independent assumption is often made in real-world applications (Pourret, et al., 2008).

Logistic regression, in general, is a regression model with a categorical dependent variable. It has been extensively used to model discrete choice problems in the transportation literature. The family includes several common model specifications, such as binary logit, multinomial logit, and nested logit (Wen and Koppelman, 2001). To allow the variation of coefficients among decision makers, mixed logit with random coefficients was introduced (Boyd and Mellman, 1980). However, it is sometimes hard for the logistic regression to capture the nonlinear and complicated influence of independent variables in the real world.

One intuitive way to handle the nonlinear problems is to employ machine learning methods, which have evolved for almost 60 years. Some typical examples of machine learning algorithms are artificial neural networks (ANNs) (McCulloch and Pitts, 1990), decision trees (Quinlan, 1986), and support vector machine (SVM) (Cortes and Vapnik, 1995). They were developed to address the classification problem and have shown superior prediction accuracies. For example, ANNs frequently outperform on huge and complex problems (Géron, 2017).

3. MOBILE DEVICE LOCATION DATA

Mobile device location data from cellphone, GPS, and location-based services (LBS) technologies have become increasingly available for transportation planning and operations. Typically, mobile device sensors record latitude and longitude coordinates, time stamps, and in some cases, additional information such as heading, positioning accuracy, acceleration rate, environmental readings, and data usage by app and time period. In practice, mobile devices have been primarily employed as probes for the estimation of aggregate travel and traffic patterns, such as OD tables, mode share, travel time, speed, traffic volume, annual average daily traffic, turning movements, and vehicle miles traveled. In this project, the research team worked with all major sources of location data: cellphone data, in-vehicle GPS data, and LBS data. **Figure 4** shows the comparison of these data sources for the city of Baltimore. In this figure, green represents higher density observations and blue represents lower density observations.



Cellphone data

In-vehicle GPS data

LBS data source 1

LBS data source 2

Figure 4. Demonstration of different sources of mobile device location data for the city of Baltimore

Table 2 presents a general comparison between the major sources of mobile device location data. It should be noted that this comparison is based on the current technologies, but the field is very dynamic and the technologies are rapidly changing. The table compares overall technologies;

specific datasets by these technologies might compare differently (**Table 3** shows how different datasets from a single technology can significantly vary in terms of data quality). In this table, "+" represents relative strength and "-" represents relative weakness. Sample size refers to the number of unique devices; observation frequency refers to the number of unique sightings per device per time-period; location accuracy represents the geospatial accuracy of records; multimodal represents the coverage of different travel modes; and truck identification represents the ease of filtering truck data from the dataset. In-vehicle GPS datasets are usually strong in terms of observation frequency and location accuracy. However, they usually suffer from small sample sizes and are limited to vehicle locations. One of the main strengths of the in-vehicle GPS data is in their ease of filtering truck data, as in-vehicle GPS data usually come with provider information, which makes it easy to filter truck locations. Cellphone data tend to benefit from a larger sample size, but they usually suffer from lower location accuracy. However, this technology is rapidly updating by the advancements in cellular technology such as 5G. Due to the strengths of LBS data with the person movements and the strengths of in-vehicle GPS data with truck movements, we selected LBS and in-vehicle GPS data for our person and truck final products, respectively.

Table 2. Comparison between major sources of mobile device location data

	Sample Size	Observation Frequency	Location Accuracy	Multimodal	Truck Identification
In-Vehicle GPS	-	+	+	-	+
Cellphone	+	-	-	+	-
LBS	+	+	+	+	-

Some common issues, such as unordered and duplicated records, need careful treatment before extracting any information from mobile device location data. The state-of-the-practice methods for raw data cleaning and quality control often include identifying and merging duplicate device observations, removing outliers, and checking on the obvious data consistency issues (e.g., devices with unreasonably high-speed readings). **Figure 5** shows a general data cleaning procedure for mobile device location data taken by the research team based on the first two out of the four dimensions of data quality assessment: consistency, accuracy, completeness, and timeliness (Batini, et al., 2009).

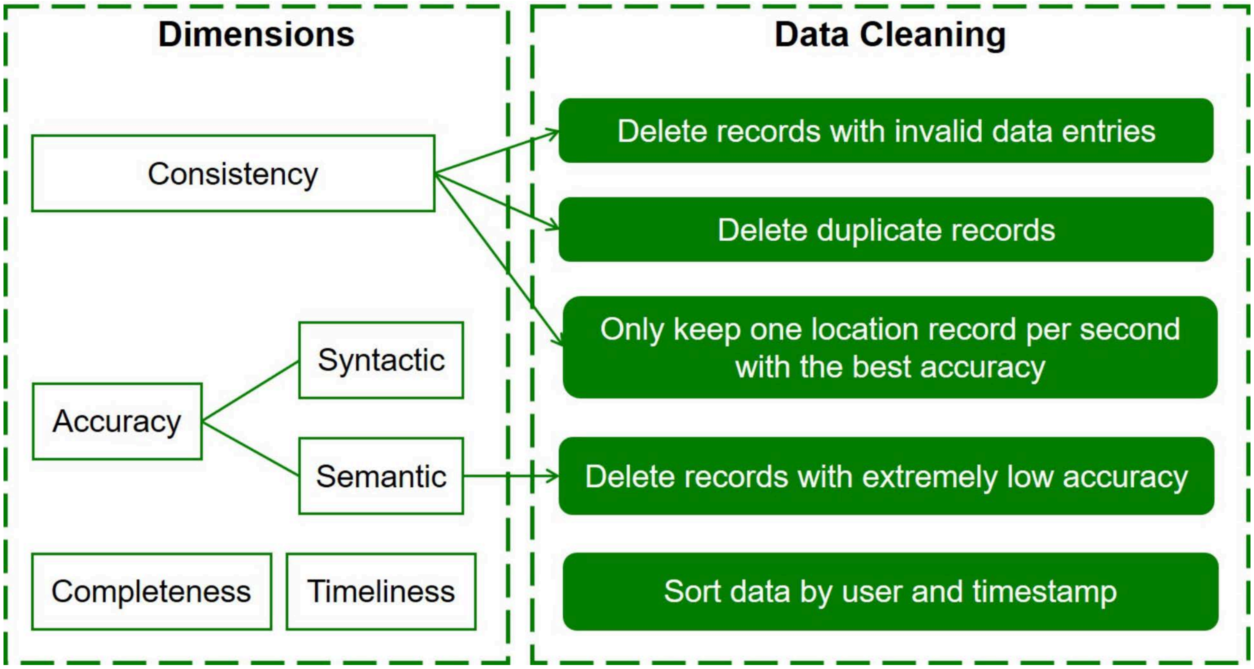

Dimensions
Data Cleaning
Consistency
Accuracy
Syntactic
Semantic
Completeness
Timeliness
Delete records with invalid data entries
Delete duplicate records
Only keep one location record per second with the best accuracy
Delete records with extremely low accuracy
Sort data by user and timestamp

Figure 5. State-of-the-practice data cleaning procedure

The completeness dimension cannot be considered without prior knowledge of the actual individual movements and mobile device usage. The timeliness is addressed by using recent mobile device datasets for any application. For the first two dimensions related to data cleaning, the consistency dimension defines certain semantic rules that a set of data items should obey. A common type of semantic rule is integrity constraints, including intra-relation and inter-relation constraints. Intra-relation constraints determine the domain of acceptable values for an attribute. For example, the latitude and longitude of a location observation should be within a reasonable range. Inter-relation constraints consider relations from different attributes. According to the integrity constraints, the cleaning procedure first deletes records with invalid entries and duplicate records to reduce redundancy. Since one subject cannot be at more than one place at the same time, the procedure keeps only one location record per second (with the highest accuracy, if applicable). Another important dimension of data quality assessment is accuracy, including syntactic and semantic accuracy. The syntactic accuracy measures the closeness of a value to all the elements of its corresponding definition domain, which is similar to the intra-relation constraints for the data. The semantic accuracy measures the closeness of a value to its real-world value. For example, an accuracy of 10 meters in a location sighting indicates that the subject should be within a radius of 10 meters from the observed location with a certain confidence level, e.g., 95%. Therefore, the cleaning procedure removes the noisy records with extremely poor accuracy, e.g., two miles.

Person location data providers describe their sample sizes with statistics such as daily active users (DAU) and monthly active users (MAU). MAUs are devices that are observed at least once a month and DAUs are devices that are continuously observed throughout the month. Reported data coverage by major data providers ranges between 5% to 70%, depending on whether they report MAU or DAU and how they define active users. While the overall sample size is measured by daily and monthly active users, these measures do not take into consideration that some devices may provide many sightings every day while other devices may only provide a few sightings in a very small number of days. In addition to applying the state-of-the-practice methods already adopted by most commercial data providers, the research team has developed a more comprehensive procedure that employs standardized quality metrics. Our algorithms not only check the consistency of each data record to ensure each raw data entry has a reasonable value, but also check to ensure that the overall distribution of data accuracy meets a certain standard. Our raw data quality metrics are listed and explained below:

- Population coverage: number of devices divided by the population of the study area.
- Temporal consistency: average number of days a device is observed in the study period.
- Frequency: the average location observations per device per day.
- Geographical representativeness: variance of population coverage among different zones of the study area, measured by Gini coefficient. Gini coefficient is a measure of equity, falling between 0 and 1, with 0 indicating equal sampling rate in all zones and 1 indicating that all observed devices are from a single zone.
- Device representativeness: a measure of the variance in the location point frequency among observed devices. This measure shows if observed devices are comparable in terms of their data frequency and are also measured by a Gini coefficient falling between 0 and 1. Raw data representativeness has a lower value if all observed devices have more consistent data frequency. Representativeness across socio-demographic groups is also important for data quality check. The same representativeness measure, when compared across socio-demographic groups, identifies socio-demographic biases that must be addressed through proper weighting.
- Hourly and daily temporal coverage: a measure of the variation of the number of location point observations among different hours of the day and different days of the month, respectively. Lower values between 0 and 1 indicate a more equitable distribution.

Table 3 presents a quality comparison among U.S. person travel raw LBS datasets from three commercial data providers. As stated before and suggested by the results, even different raw data sources from the same technology show significant quality differences.

Table 3. Data quality comparison among three commercial LBS datasets

Selected Raw Data Quality Metrics	**Dataset 1**	**Dataset 2**	**Dataset 3**
Population Coverage (%)	5.85	36.5 (best)	23.92
Geographical Representativeness (0~1)	0.13	0.09 (best)	0.12
Frequency (observations per device per day)	57	75	190 (best)
Temporal Consistency (days per device)	10.18	12.90	14.67 (best)
Device Representativeness (0~1)	0.71	0.67 (best)	0.81
Hourly Temporal Coverage (0~1)	0.67	0.64	0.249 (best)
Daily Temporal Coverage (0~1)	0.24	0.05	0.03 (best)

Mobile device data for trucks usually come from in-truck GPS devices. Since truck data providers aggregate observations from many original sources, the resulting truck raw datasets include location data from a variety of vehicles such as vehicle service fleets, field service and local delivery fleets, for-hire and private trucking fleets, truck fleets from larger corporations, and even some consumer vehicles and taxi fleets that should not have been included. Similarly, the research team have developed raw data cleaning procedures and truck data quality metrics to ensure the consistency, accuracy, completeness, representativeness, and timeliness of the truck raw data.

4. METHODOLOGY

4.1. Overall Framework

Figure 6 shows the steps required to produce person OD tables from the anonymized raw data. Open-source algorithms with aggregate- and individual-level validation were developed for each one of the computation steps.

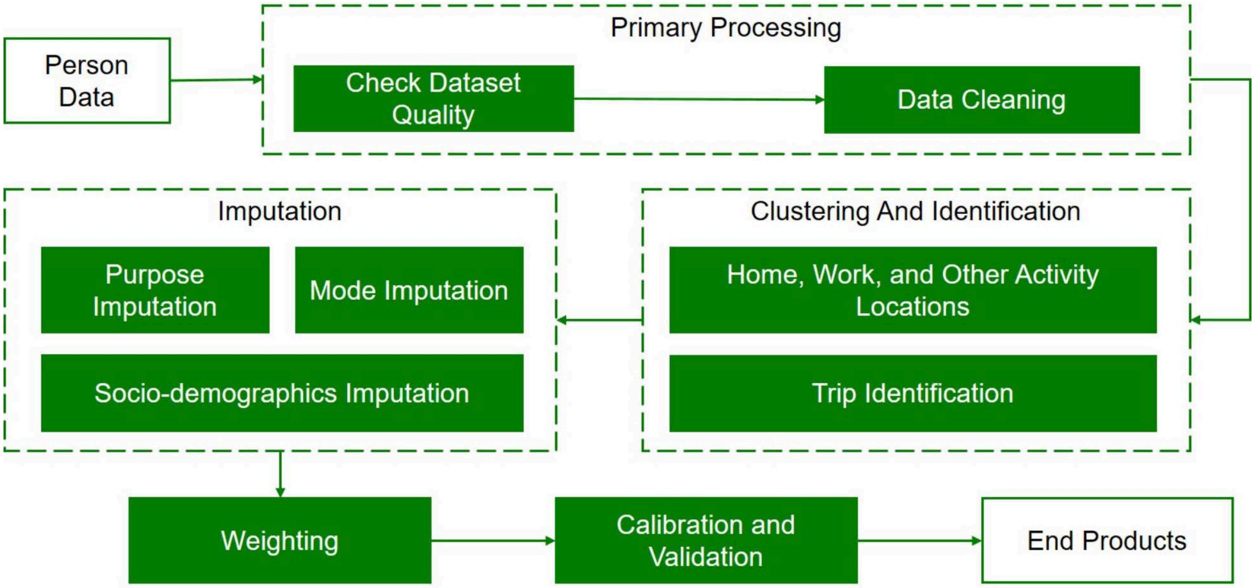


Figure 6. Person OD production methodology

Data cleaning: Data come from various sources, varying in size, frequency, accuracy, etc. We developed a set of standards and rules to check overall dataset quality and clean the data from low-quality observations.

Data clustering: The original datasets include anonymized location observations, without any information about activities or trips. At this step, locations from each device are aggregated to activity points through density-based clustering and home and work clusters (for workers with a fixed work location) are identified.

Trip identification: Trips are the unit of analysis in OD production, however, the raw data does not identify which points form a trip together. In this step, trips are identified through a set of rules using identified activity clusters.

Imputation: The original data lack useful information such as travel mode, trip purpose, and traveler socio-demographic information. These are important attributes for any planning application or policy analysis. We developed open-source statistical and machine learning algorithms to impute these non-observable attributes.

Sample normalization, weighting, and expansion: Even though mobile device location datasets benefit from a large sample size, they still do not cover the entire population and suffer from several sources of bias. The sample coverage varies among different time periods, different population groups, and different geographies. Using census information, imputed socio-demographics, and market penetration information, we have developed methods to normalize and expand the data sample to produce population OD information.

Calibration and validation: We have validated the research products at both individual algorithm level and aggregate result level through comparison with a most comprehensive set of validation datasets. In some cases, we calibrated the imputation model parameters to match the validation goals.

The procedure for producing truck OD tables (**Figure 7**) follows similar steps, except that no imputation is needed. The weighting procedure is also different for truck data, as the set of information used for weighting person data, such as census population information, are not available for truck data. Truck weighting depend heavily on traffic count data.

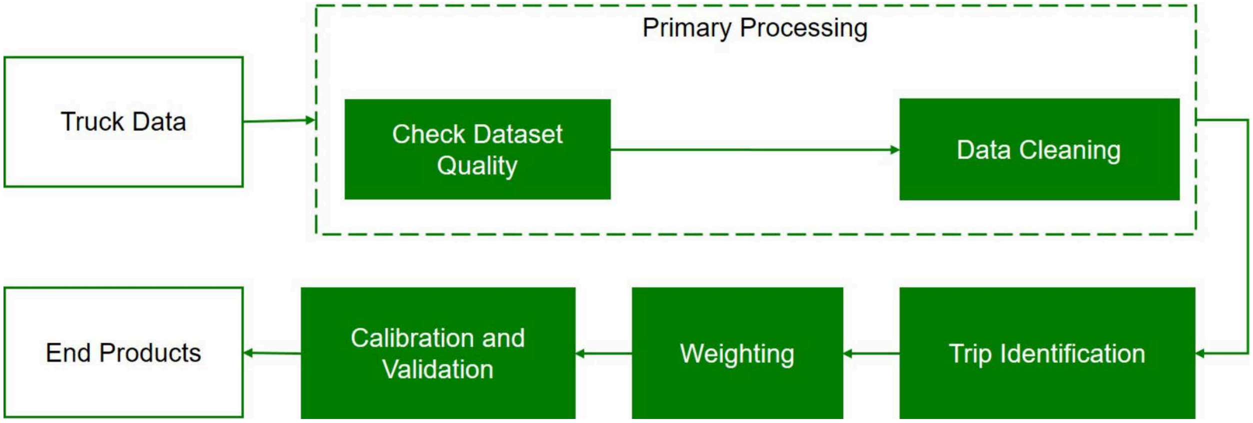


Figure 7. Truck OD production methodology

4.2. Trip Identification

Trips are the unit of analysis for almost all transportation applications. Traditional data sources such as travel surveys include trip-level information. The mobile device location data, on the other hand, do not directly include trip information. Location sightings can be continuously recorded while a device moves, stops, stays static, or starts a new trip. These changes in status are not recorded in the raw data. As a result, researchers must rely on trip identification algorithms to extract trip information from the raw data. Basically, researchers must identify which locations form a trip together. The state-of-the-practice methods to identify trips from raw location data points are as follows:

- Method 1: Consider the time and distance relationship between consecutive location point observations to identify moving points and static points. Consecutive moving points between two sets of static points form a trip. The limitation of this approach is that any static point, such as a stop at a red light, can be identified as a trip end. Also, many short trips are formed due to local moves that do not represent any real trip.
- Method 2: Consider zone boundaries to identify movements from one zone to another, which is applicable to identifying inter-zonal trips only. The limitation of this approach is

its inability to model internal trips. It can also wrongly merge separate trips together or break a single trip into separate trips.

- Method 3: Identify location point clusters as activity locations with spatial clustering methods. Location point observations between two consecutive activity locations form a trip. The limitation of this approach is that not all trip ends are necessarily identified as activity clusters, so many trips may be missed.

The research team developed a new trip identification algorithm that utilizes both cluster information and moving/static patterns (Methods 1 and 3). The algorithm runs on the observations of each device separately. The following sections summarize the algorithm steps.

4.2.1. Activity Clustering

The algorithm starts by clustering all device observations into activity locations using HDBSCAN (Ester, et al., 1996) clustering algorithm. This step takes the cleaned multi-day location data as input and applies an iterative algorithm until no cluster has a radius larger than two miles. The iterative algorithm consists of two parts: HDBSCAN based on a minimum number of point parameters and filtering non-static clusters based on time and speed checks. After finalizing the potential stay clusters, the framework combines nearby clusters to avoid splitting a single activity (**Figure 8**). The identified activity clusters are used in trip identification, home location identification, and work location identification.

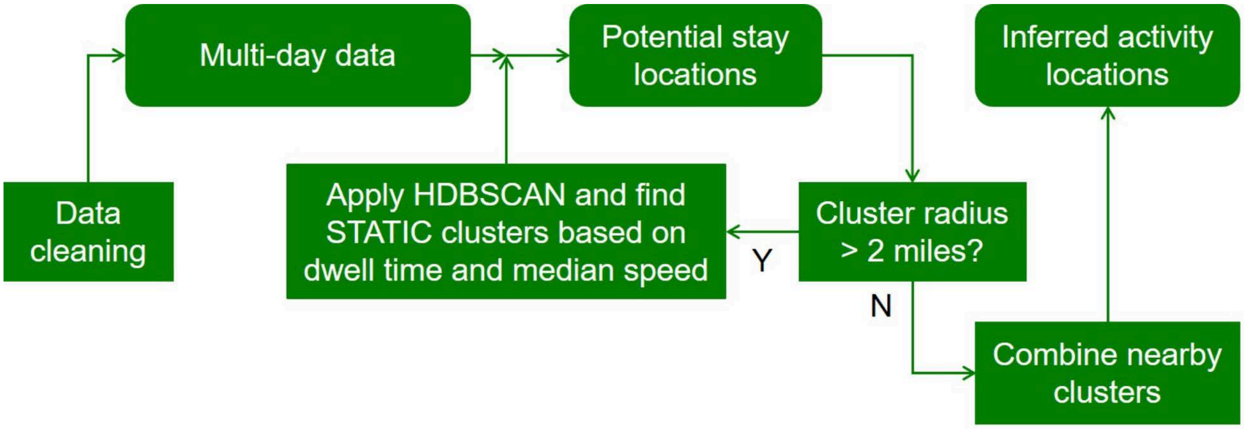

Multi-day data
Potential stay locations
Inferred activity locations
Data cleaning
Apply HDBSCAN and find STATIC clusters based on dwell time and median speed
Y
Cluster radius > 2 miles?
N
Combine nearby clusters

Figure 8. Activity clustering methodology

4.2.2. Pre-Processing

First, all device observations are sorted by time. The trip identification algorithm assigns a hashed ID to every trip it identifies. The location dataset may include many points that do not belong to any trips. The algorithm assigns “0” as the trip ID to these points to identify them as static points. In this pre-processing step, for every visit of the static clusters, the trip ID of all observations besides the first and the last observation is set to “0”. The first observation may belong to the trip “to” the static cluster and the last observation may belong to the trip “from” the static cluster. Next, for every observation, the following attributes are calculated:

- Distance to: distance from the previous observation to the current observation (0 for the first observation of the device).
- Distance from: distance from the current observation to the next observation (0 for the last observation of the device).
- Time to: time from the previous observation to the current observation (0 for the first observation of the device).
- Time from: time from the current observation to the next observation (0 for the last observation of the device).
- Speed to: speed from the previous observation to the current observation (0 for the first observation of the device).
- Speed from: speed from the current observation to the next observation (0 for the last observation of the device).

The trip identification algorithm has three hyper-parameters: distance threshold, time threshold, and speed threshold. The speed threshold is used to identify if an observation is recorded on the move. The distance and time thresholds are used to identify trip ends. At this step, the algorithm identifies the device's first observation with $speed\ from \geq speed\ threshold$. This identified point is on the move, so a hashed trip ID is generated and assigned to this point. All points recorded before this point, if they exist, are set to have "0" as their trip ID. Next, the recursive algorithm identifies if the next points are on the same trip and should have the same trip ID.

4.2.3. Recursive Algorithm

This algorithm (**Figure 9**) checks every point to identify if they belong to the same trip as their previous point. If they do, they are assigned the same trip ID. If they do not, they are either assigned a new hashed trip id (when their $speed\ from \geq speed\ threshold$) or their trip ID is set to "0" (when their $speed\ from < speed\ threshold$). Identifying if a point belongs to the same trip as its previous point is based on the point's "speed to", "distance to" and "time to" attributes. If a device is seen in a point with $distance\ to \geq distance\ threshold$ but is not observed to move there ($speed\ to < speed\ threshold$), the point does not belong to the same trip as its previous point.

When the device is on the move at a point ($speed\ to \geq speed\ threshold$), the point belongs to the same trip as its previous point; but when the device stops, the algorithm checks the radius and dwell time to identify if the previous trip has ended. If the device stays at the stop (points should be closer than the distance threshold) for a period of time shorter than the time threshold, the points still belong to the previous trip. When the dwell time reaches above the time threshold, the trip ends, and the next points no longer belong to the same trip. The algorithm does this by updating "time from" to be measured from the first observation in the stop, not the point's previous point.

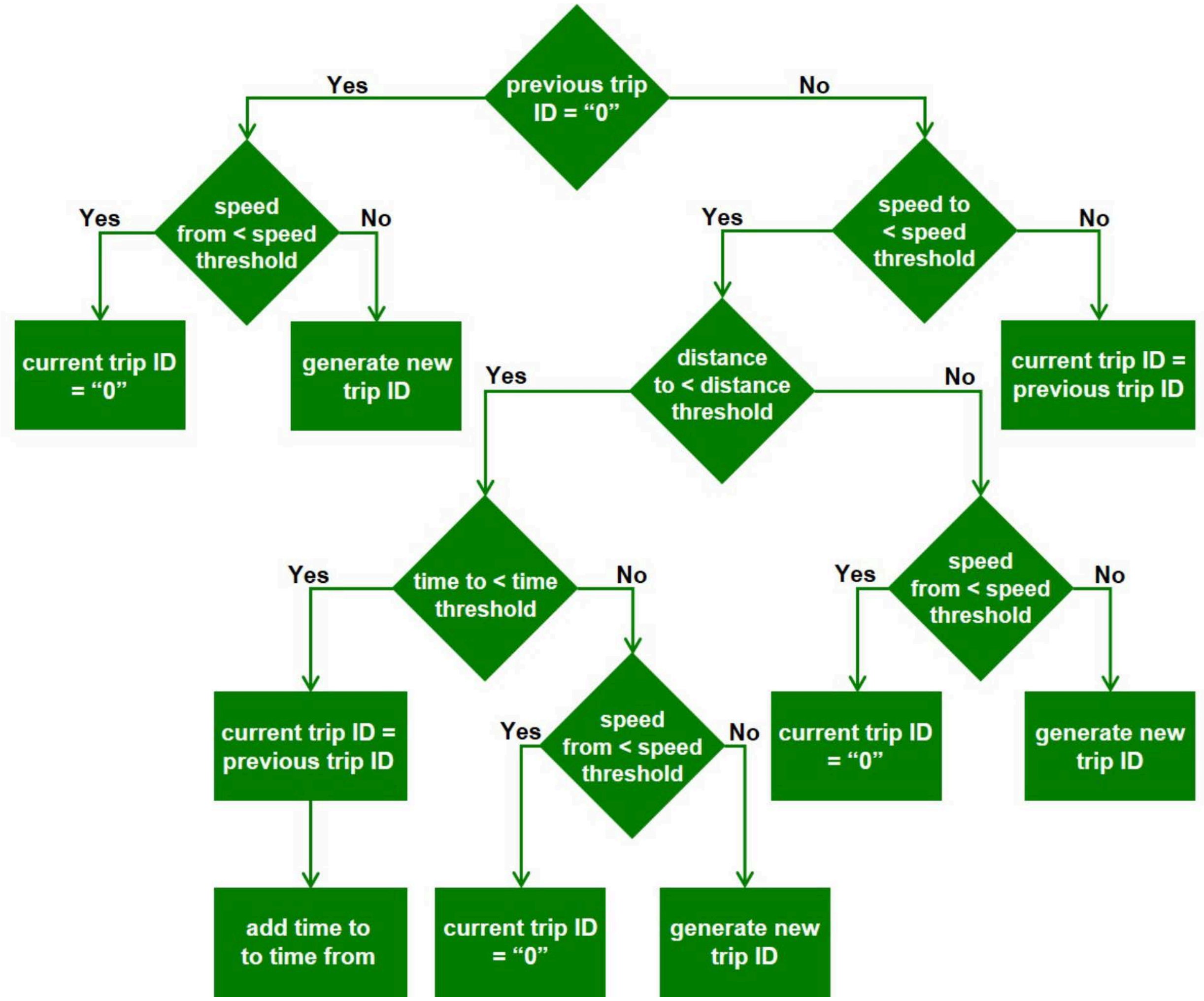


Figure 9. Trip identification methodology

4.2.4. Post-Processing

The algorithm may identify a local movement as a trip if the device moves within a stay location. To filter out such trips, all trips that are within a static cluster and all trips that are shorter than 300 meters are removed.

4.2.5. Validation

Figure 10 and **Figure 11** show the validation of this algorithm by running the algorithm on a sample of national mobile device location data and comparing the trip lengths and travel times with the reported travel distances and travel times from the NHTS 2017. A satisfactory match is observed between the two datasets, except that our algorithm shows more longer-duration trips. We need to note that this algorithm validation includes all trips, not just long-distance trips. Product validation, for both national and MPO products, is presented later in the **Section 5**.

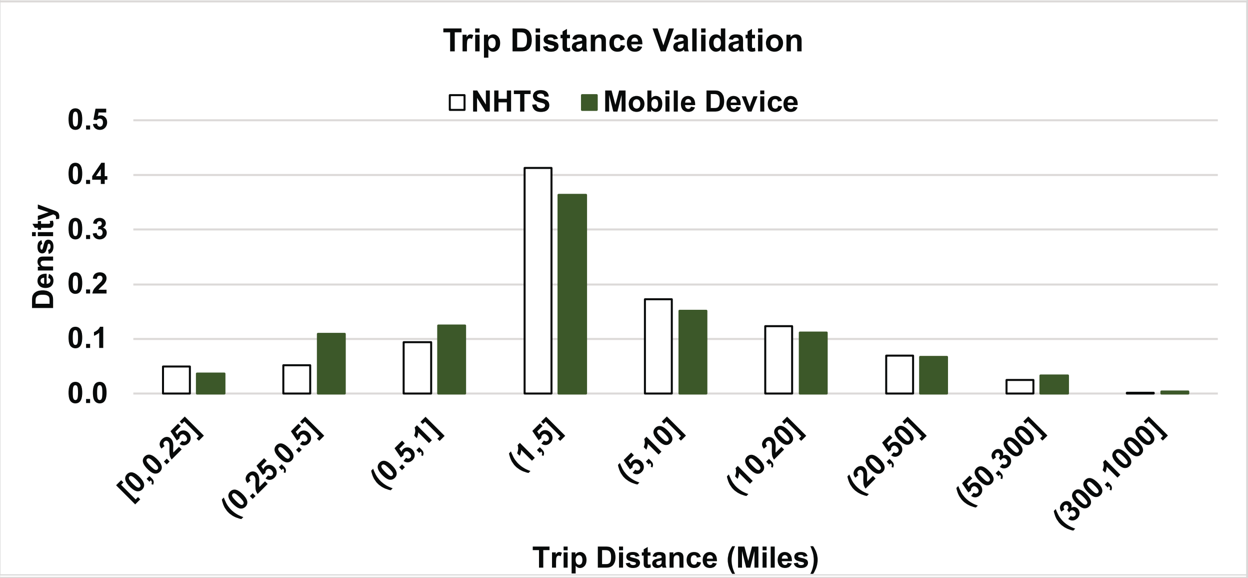


Figure 10. Distance validation of the trip identification algorithm against NHTS2017

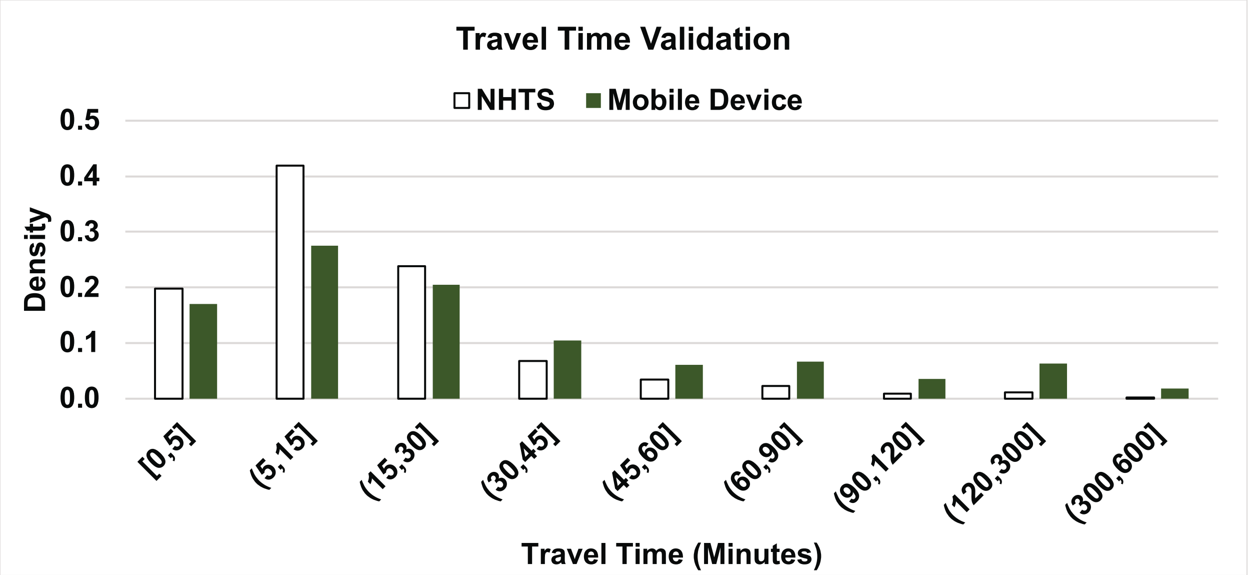


Figure 11. Travel time validation of the trip identification algorithm against NHTS2017

Trip identification for truck trips is significantly simpler. The continuous frequent sightings in the GPS datasets make it possible to implement a simpler trip identification algorithm. For truck trip identification, truck trips are formed by grouping the sequences of GPS points from the same truck GPS device when the device is moving. A truck device is considered moving if the device progresses at least 200 meters within 10 minutes. As long as the threshold is met, a truck trip continues.

4.3. Travel Mode Imputation

Travel mode detection based on mobile device location data has garnered increased research attention in the past decade. Researchers have explored various artificial intelligence (AI) methods to cope with the travel mode detection problem, including decision trees, neural networks, naïve Bayes and Bayesian networks, support vector machines, random forests, etc. Overall, the state-of-the-art in research works can detect drive mode with high accuracy (utilizing high-resolution location traces, which also draw battery concerns); however, the detection of bus and rail modes is not as satisfying. One possible methodological limitation that could lead to this is the single-layer AI representation. The single-layer neurons or rules often cannot handle a high-dimensional problem. To generalize the unobserved feature combinations for better detection of bus and metro modes, a multi-layer deep neural network (DNN) can be used, which performs efficiently with much fewer nodes in each layer. Despite some advancement in the state-of-the-art, the state-of-the-practice is mainly limited to the identification of non-motorized mode based on speed and some acceleration information that are not always available.

Our research team developed a jointly trained single-layer model and deep neural network for travel mode detection of this project. This model combines the advantages of both types of models to be able to make sufficient generalizations using a multi-layer DNN and capture the exceptions using the wide single-layer model. The datasets used for training the model were collected from the incenTrip mobile phone app, developed by MTI, where the ground truth information for car, bus, rail, bike, walk, and air trips was collected. To effectively detect the travel mode for each trip, feature construction is critical in providing useful information. Travel mode-specific knowledge is needed to improve the detection accuracy. In addition to the traditional features used in the literature (e.g. average speed, maximum speed, trip distance, etc.), we also integrated the multi-modal transportation network data to construct innovative features in order to improve the detection accuracy based on network data integration (**Table 4**).

Table 4. Mode imputation features

Features	Unit
Sample Rate Feature	
Average # of records per minutes	number / minute
Trip Features	
Trip distance	meters
Origin-destination distance	meters
Trip time	minutes
Average speed	meters
Minimum speed	meters
Maximum speed	meters
Median speed	meters
5-percentile speed	meters
25-percentile speed	meters
75-percentile speed	meters
95-percentile speed	meters
Multimodal Transportation Network Features	
Min, max, median, 5-, 25-, 75-, 95-percentile distance to the rail network	meters
Min, max, median, 5-, 25-, 75-, 95-percentile distance to the bus network	meters
Min, max, median, 5-, 25-, 75-, 95-percentile distance to the drive network	meters

Figure 12 shows the prediction accuracy of our method (wide and deep learning on the right) and compares it with the other methods we tested. The wide and deep learning method can achieve over 95% prediction accuracy for drive, rail, air, and non-motorized, and over 90% for bus modes.

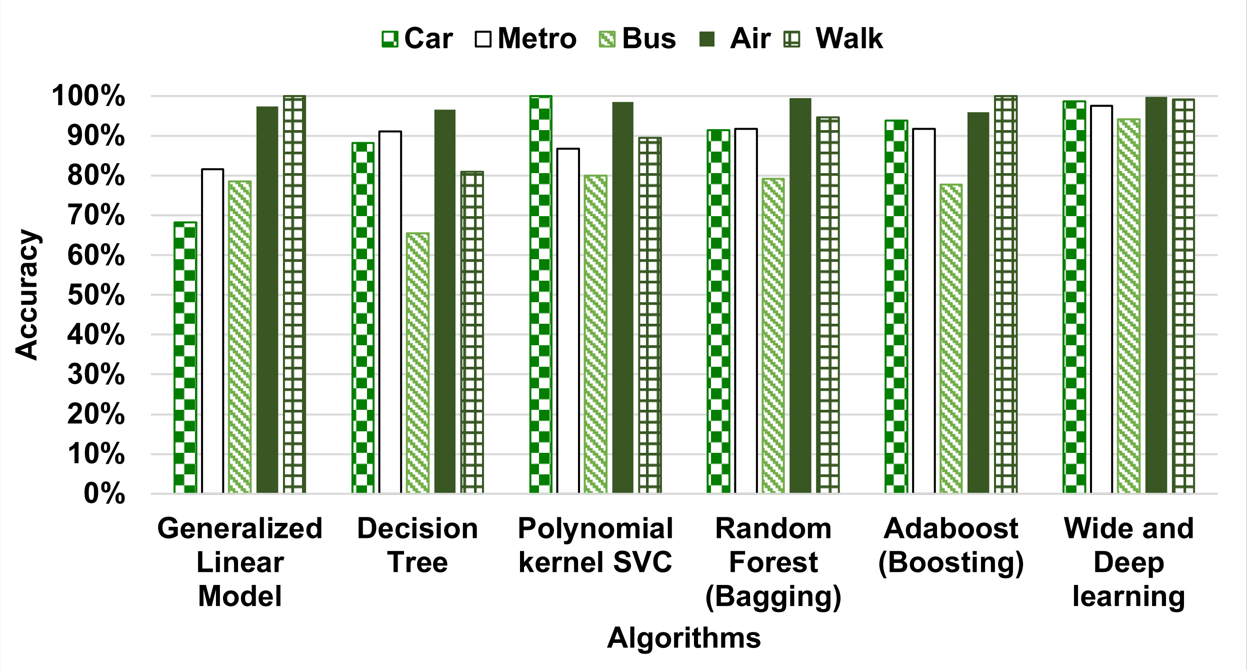


Figure 12. Comparison of different tested mode imputation algorithms accuracy at individual level

The research team successfully applied the trained algorithms to develop multimodal person travel OD tables for the project (see **Figure 13** that shows raw location data points by different travel modes).

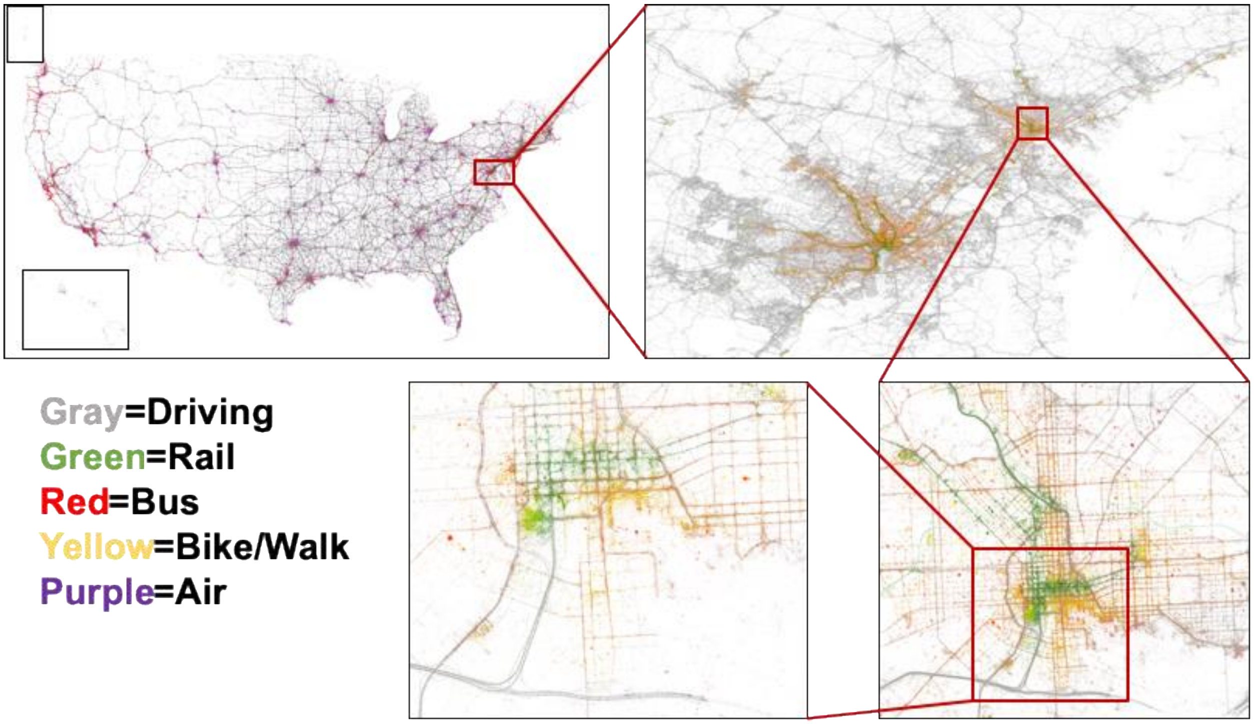


Figure 13. Demonstration of the multi-modal travel patterns

4.4. Trip Purpose Imputation

The identification of activity purposes can directly determine the trip purpose. In addition, activity purposes along with activity locations help to extract the socio-demographic information. Such imputed information contributes to the sample weighting and extrapolation. The activity clustering was explained in 4.2.1. Here, we explain how the activity purposes can be identified and how the identified activity purposes inform trip purpose imputation.

Our research team developed a method that considers the behavior and land use features to identify home and work locations and to make the type of evaluation of other activity clusters based on the point of interest (POI) data. **Figure 14** shows the methodology that focuses on the home and work location identification since the two location types cannot be solely inferred from the POI data with confidence. Instead of setting a fixed time period for each type, e.g., 8pm to 8am as the study period for home location identification and the other half day for work location identification, the framework examines both temporal and spatial features for the entire activity location list. The benefits are two-fold: the results for workers with flexible or opposite work schedules would be more accurate and the employment type for each device could be detected simultaneously. We must note that school will also be identified as a work location, so school trips are categorized as work trips. **Figure 15** shows the validation of home and work location imputations by comparing the distance from home to work between longitudinal employer-household dynamics (LEHD) data and the imputed locations for a set of mobile device location data for the Baltimore metropolitan area. We can observe a satisfactory match. We must note that this algorithm validation includes all trips. Product level validation for both national and MPO products is presented later in the **Section 5.**

The trip purpose identification for MPO Person OD products is based on the imputed activity purposes. The primary purposes are home-based work (HBW), home-based other (HBO), and non-home based (NHB) trips, which are widely used in the state of the practice.

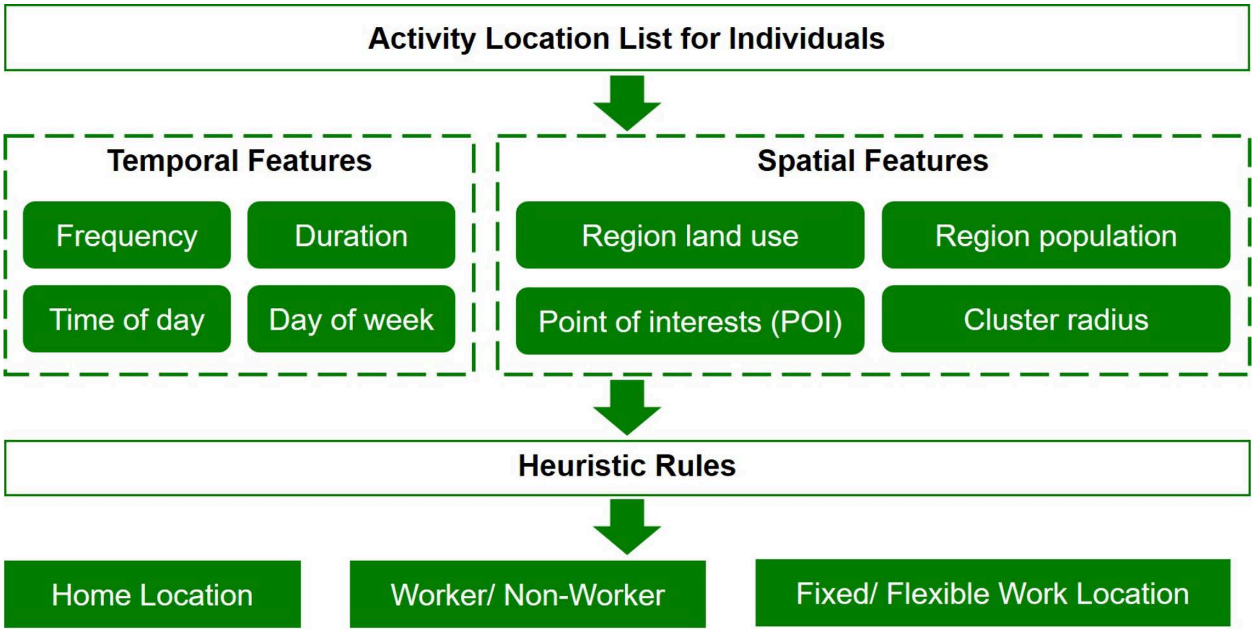


Figure 14. Activity purpose imputation methodology

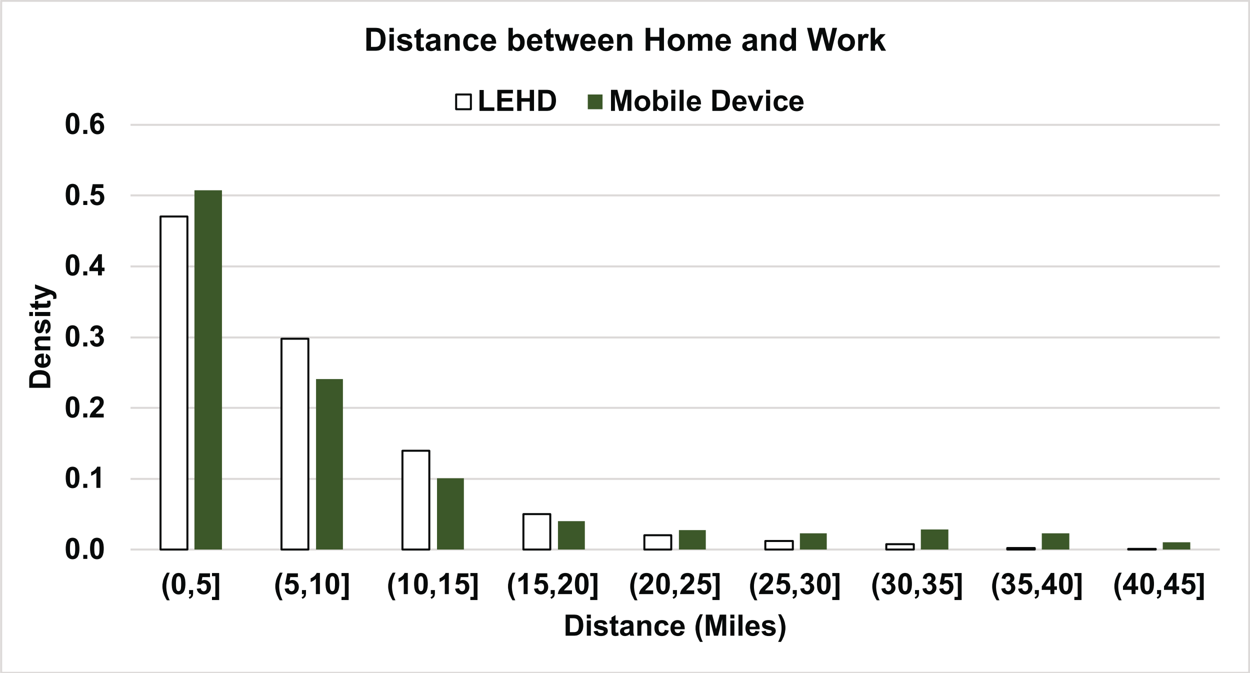


Figure 15. Validation of home and work imputation against LEHD

In addition to local daily trips, long-distance trips – either periodic or not – are of importance for many planning applications. The national OD tables in this project only include long-distance trips longer than 50 miles, for which typical daily trip purposes are not applicable. Capturing long-distance travel has always been a challenge for traditional household travel surveys. Our research team has developed a separate trip purpose imputation model to be applied to the long-distance trips, with the following trip purpose categories: business, personal business, and pleasure. **Figure 16** shows the modeling framework. We have specifically developed an imputation model for long-distance tours. The model was trained and verified by comparisons with the American Travel Survey (ATS) and recentzl surveys.

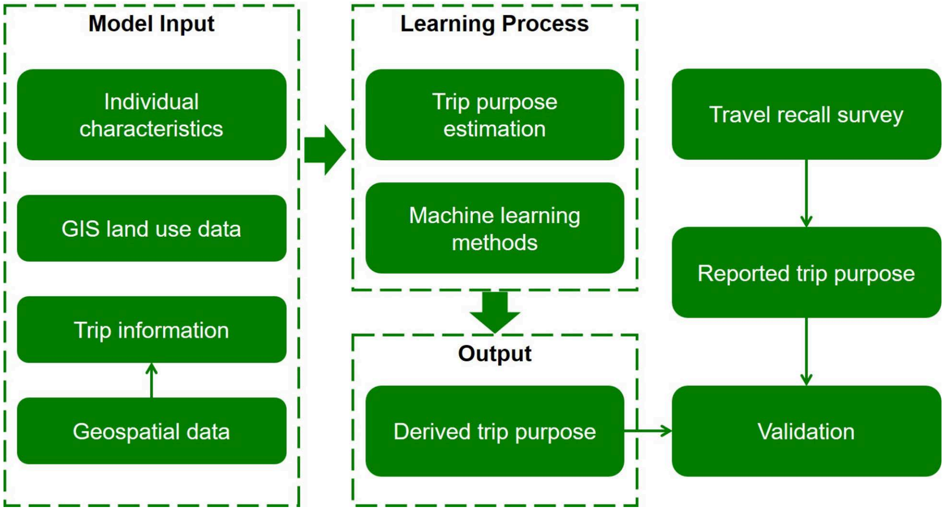
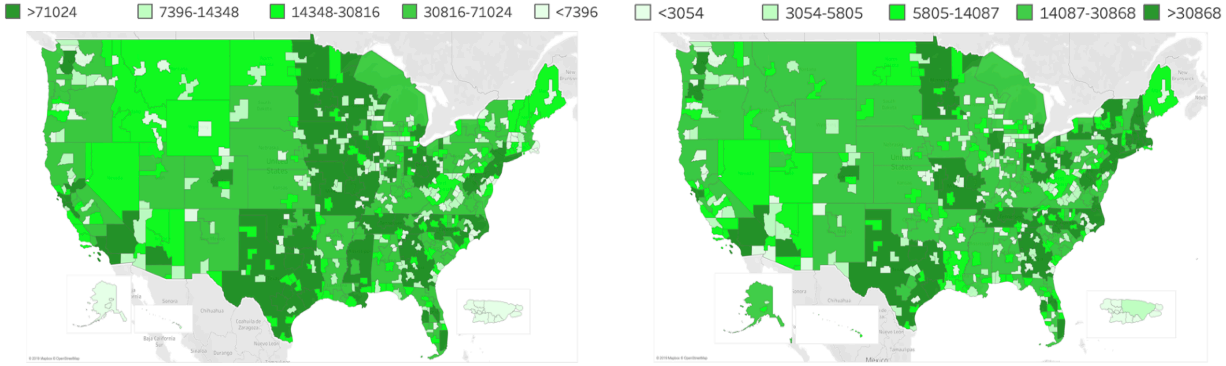


Figure 16. Long-distance purpose imputation methodology

The long-distance trip purpose imputation method was applied to the national mobile device location data to produce national OD separated by long-distance purpose categories. **Figure 17** is an example of the national results. Figure (a) shows the comparison of total business trips on an average weekday in the first quarter of 2017 among different geographies. Figure (b) shows the comparison of total pleasure trips on an average weekend in the second quarter of 2017 among different geographies.



(a) Number of business trips generated from each zone on an average weekday in quarter 1

(b) Number of pleasure trips attracted to each zone on an average weekday in quarter 2

Figure 17. Imputed trip purpose for a set of national mobile device location data

4.5. Socio-Demographic Imputation

Due to privacy concerns, mobile device location data contain very little ground truth information about the device owners. However, it is essential to understand how representative the sample is and how different segments of the population travel. The state-of-the-practice method is to assign either the census population socio-demographic distribution or the public use microdata sample (PUMS) units to the sample devices within the same geographic area based on the imputed home locations. Such imputation methods are unable to identify or address some of the potential sample biases in the mobile device data. As the last imputation task in the framework, socio-demographic imputation can exploit all the activity-level and trip-level information. **Figure 18** shows the socio-demographic imputation methodology developed by the research team that considers both travel behavior and environmental features. The socio-demographic imputation happens at the device level, applied for both national and MPO person OD products.

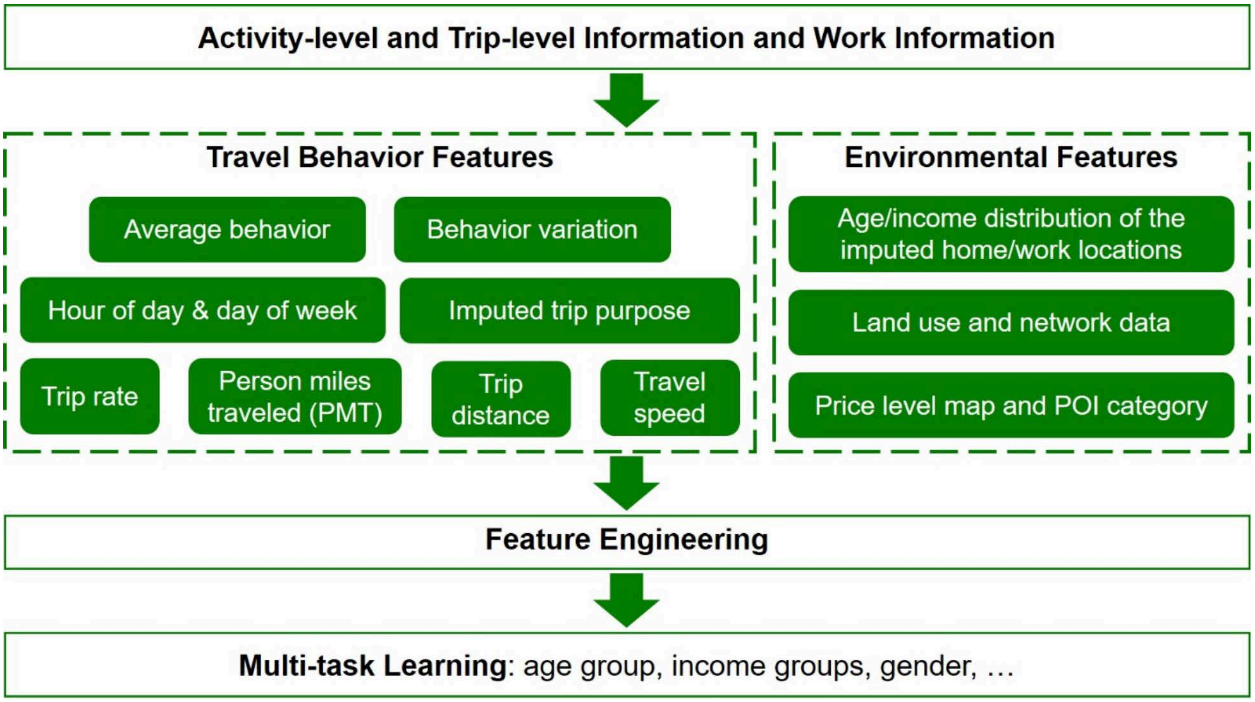
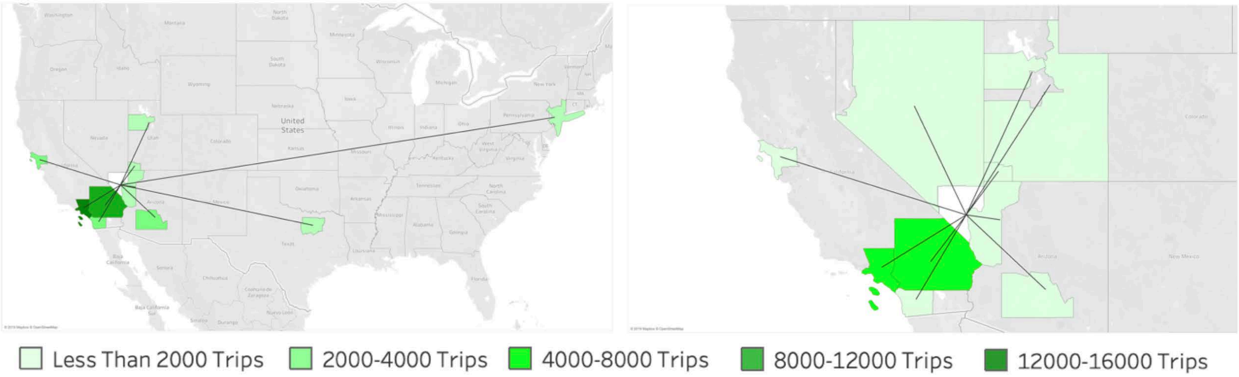

Activity-level and Trip-level Information and Work Information
Travel Behavior Features
Average behavior
Behavior variation
Hour of day & day of week
Imputed trip purpose
Trip rate
Person miles traveled (PMT)
Trip distance
Travel speed
Environmental Features
Age/income distribution of the imputed home/work locations
Land use and network data
Price level map and POI category
Feature Engineering
Multi-task Learning: age group, income groups, gender, …

Figure 18. Socio-demographic imputation framework

For travel behavior features, the method considers both average behavior and behavior variation in different time-of-day and day-of-week for different purposes. To quantify the travel behaviors, the method includes features like trip rate, trip distance, travel speed, etc. For environmental features, it is intuitive to evaluate the socio-demographics distribution for the imputed home and work location from census ACS data. In addition, the land use data, e.g., areas for various industry sectors and transportation network density, indicating the underlying reasons for home location selection may be distinctive for different socio-demographic groups. Another type of helpful environmental features is the frequently mentioned POI type. Even more helpful is the price level for establishments.

The socio-demographic imputation requires a significant amount of computation, as various features from different databases should be calculated and used. In order to balance the computations of this project, the current deliverables are based on the state-of-the-practice method. The state-of-the-art method is further described in **Section 6.2**, as it is continuously being improved by adding new features and fine-tuning the model configuration. Figure 19 is an example of applying the model over a national mobile device location dataset. Figure (a) shows the top ten destinations of the high-income devices for trips generated from the Las Vegas MSA and Figure (b) shows the top ten destinations of the low-income devices for trips generated from the Las Vegas MSA.



(a) High- income

(b) Low- income

Figure 19. Top destinations for trips originated from Las Vegas MSA by income group

4.6. Weighting

Despite the advancements in statistical methods used for weighting the core survey observations, the state-of-the-practice in weighting mobile device location observations has not yet been reliable enough. Most of the available products from mobile device location data are either unweighted or not weighted based on a comprehensive framework. There are three popular weighting methods that have been widely used: 1) estimate device-level weights based on a single factor such as the overall penetration, 2) estimate device-level weights based on sample coverage in the imputed home location, and 3) estimate trip-level weights by calibrating the products with traffic counts without specific treatments to the device-level weights. Such weighting methods are unable to identify or address the potential mobile device sample biases, as different socio-demographic groups may have different travel behavior. **Table 5** illustrates the effects of two weighting methods (weighting based on sampling rate in the home location versus weighting based on the socio-demographic information) on the annual total vehicle miles traveled (VMT) and person miles traveled (PMT) estimates using a 2017 NHTS sample, where the sampling rate-based method overestimates both numbers by over 35%. Weighting based on sampling rate can produce misleading estimates for critical measures such as VMT. This highlights the importance of imputing socio-demographics and incorporating the imputed information in weighting.

Table 5. Effect of considering socio-demographics in weighting on VMT and PMT

Data Source: 2017 NHTS	Weighted by Household and Individual Social Demographics[1]	Weighted by County-Level Sampling Rate[2]	Percentage Difference
Annual VMT Estimates (millions)	2,105,882	2,925,670	+38.9%
Annual PMT Estimates (millions)	3,970,287	5,458,773	+37.5%

Although the device owners of mobile device location data are anonymous, this comparison highlights the importance of imputing the individual-level socio-demographics and incorporating such information in weighting. Such weighting also addresses the issue of representativeness and demographic biases in the mobile device location data. In addition, the socio-demographic imputation would allow stakeholders to evaluate the traffic demand and discover travel patterns (in the format of passenger OD tables) by different socio-demographic groups.

The device sample must be expanded to produce population-level statistics. The devices available in our dataset represent a sample of the population, so device-level weights are needed to expand the device sample. The research team has developed a weighting methodology based on the iterative proportional fitting (IPF) optimization method to derive device-level weights based on the imputed device-level socio-demographic information. The framework utilizes the imputed socio-demographic information and assigns weights to the devices in order to match socio-demographic distributions at the census block group for the imputed home census block group based on the Five-year American Community Survey (ACS) estimates for 2014 to 2018 from the U.S. Census Bureau.

While application needs require population level information about truck travel, properly weighted truck products produced from mobile device location data are scarce. Truck travel data are usually collected through in-vehicle GPS units; therefore, they show significant differences with the person travel data. As a result, typical weighting methods used for weighting person travel data based on home location and market penetration cannot be applied for truck travel. Our team has developed a weighting method that utilizes all available sources of truck travel information to properly weight the truck OD tables. We first run a map matching process to match the truck points to the road network. To construct the truck count data for weighting, we utilize the HPMS data for the truck AADT estimates and the TMAS data for the monthly and weight class variations in truck traffic. **Figure 20** shows the temporal variation of truck travel in different states through the ratio of maximum monthly factor over the minimum monthly factor for trucks traveling in the state, calculated using TMAS data. The observed variation highlights the importance of taking monthly variation into account. The unweighted trip roster is then divided into two subsets: one with trips passing at least one count station and the other with trips passing zero count stations. For trips passing at least one count station, an iterative weighting method is applied to derive the trip-level

[1] The estimates weighted by individual-level social demographics are computed by Westat for Federal Highway Administration.

[2] The estimates weighted by sampling rate at a county level are computed based on the geocoded location of the sampling units. The number of sampling units is summarized for each county in the U.S. Then the personal weight for each sampling unit in a county is computed by dividing the number of sampling units by the total population. The trip-level weight is computed by multiplying the personal weight by 365 as described in the 2017 NHTS Weighting Report.

weights while minimizing the overall differences between the observed and weighted truck traffic. To address the multiple-solution problem, we consult the freight analysis framework (FAF) OD data to constrain the solutions and select the optimal one. Trip-level weights for observed truck trips that pass no truck traffic count stations are derived as the average trip-level weight of all truck trips that pass count stations and have similar OD patterns in the same month.

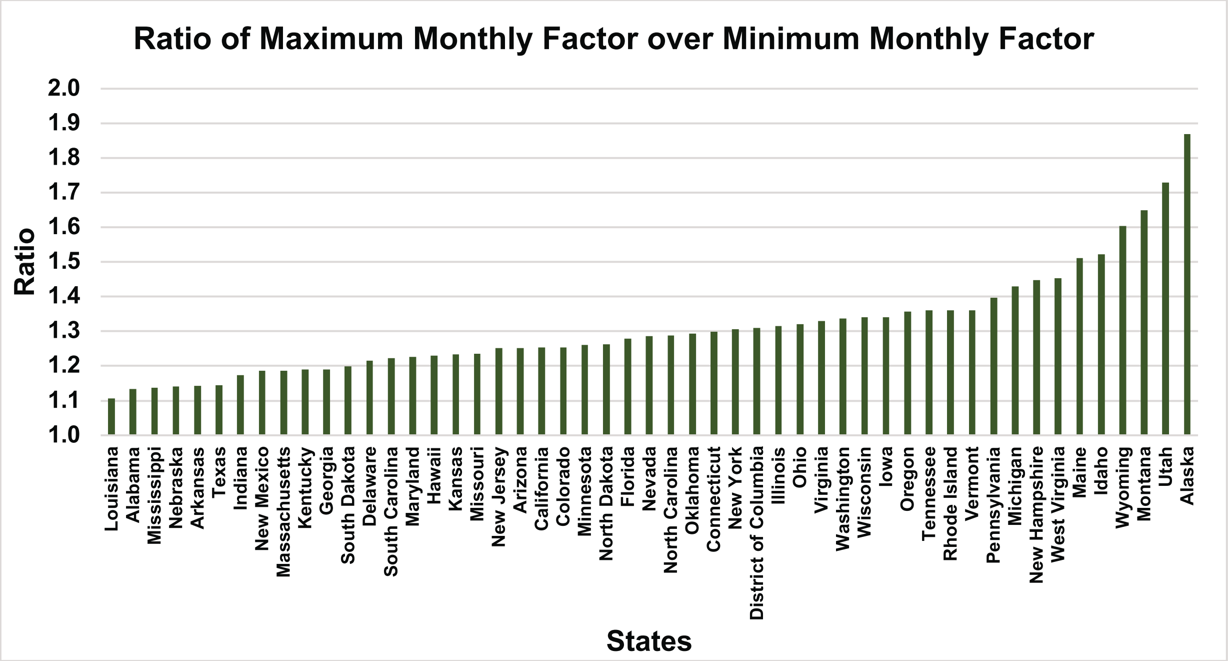


Figure 20. Monthly variation of truck trips in different states obtained from TMAS data

4.7. Validation Methodology

Validation is very critical for products produced from mobile device location data, as there are various procedures and algorithms involved in going from the raw data to the end products. Transparency and openness in validation for both algorithms and end products are required in order to build more confidence into the mobile device location data and their derived products.

The research team made every effort to produce transparent and accurate OD products that are validated at both algorithm level and product level. The team collected a comprehensive set of validation data, including smartphone survey ground-truth data collected through MTI's smartphone app (incenTrip) and public domain datasets such as travel surveys, DB1B and T100, rail and transit ridership datasets, a traffic count dataset, etc. The team made sure all products are consistent with these datasets. Overall, the project products showed satisfactory validation results (**Table 6**), which are presented in detail in the following section.

Table 6. Overall data quality for the EAR OD products

Statistic	Validation Dataset	Percent Difference
Total Number of Trips	NHTS	8.8%
Total Air Trips	DB1B	1.7%
Total Air Trips	T100	9.5%
Baltimore Rail Ridership	NTD	11.2%

5. PRODUCT DESCRIPTION

5.1. National Person Trip Product

This product contains U.S. national long-distance person travel OD data for the year 2017. The product includes annual daily average values for person travel. Each row of the CSV file corresponds to weighted person trip counts between one specific OD pair. The count information is cross-tabulated by three day types, i.e., Average_Day (Monday through Sunday), Average_Weekday (Monday through Friday) and Average_Weekend (Saturday and Sunday). **Table 7** shows the variables in the product with their description and variable type. The seventeen count variables in this table are all presented for each day type category, leading to 51 count columns in the product.

Table 7. National person OD product codebook

Name	Label	Type
Origin_Zone_ID	GeoID code of the origin zone based on the national zone system.	Numeric
Origin_Zone_Name	Name of the origin zone	Character
Destination_Zone_ID	GeoID code of the destination zone based on the national zone system.	Numeric
Destination_Zone_Name	Name of the destination zone	Character
Count	Total trip counts	Numeric
Income_lessthan_20k	Trip counts for travelers that have a household income of less than $20,000	Numeric
Income_20k_50k	Trip counts for travelers that have a household income of $20,000 to $50,000	Numeric
Income_50k_100k	Trip counts for travelers that have a household income of $50,000 to $100,000	Numeric
Income_morethan_100k	Trip counts for travelers that have a household income of more than $100,000	Numeric
Age_lessthan_35	Trip counts for travelers with age less than 35	Numeric
Age_35_65	Trip counts for travelers with age between 35 to 65	Numeric
Age_morethan_65	Trip counts for travelers with age more than 65	Numeric
Male	Trip counts of male travelers	Numeric
Female	Trip counts of female travelers	Numeric
Car	Trip counts with car as a primary mode of transportation	Numeric
Rail	Trip counts with rail as a primary mode of transportation	Numeric
Bus	Trip counts with bus as a primary mode of transportation	Numeric
Air	Trip counts with air as a primary mode of transportation	Numeric
Business	Trip counts for business trips	Numeric
Personal_Business	Trip counts for personal business trips	Numeric
Pleasure	Trip counts for pleasure trips	Numeric

Table 8 shows the comparison of total weekday daily trips between the national person OD products and the NHTS 2017. This project focused on trips longer than 50 miles in airline distance.

NHTS distances are based on network distance. As a result, mobile device trip totals are between NHTS trip totals for trips longer than 50 miles and trips longer than 75 miles.

Table 8. Comparison of person weekday trip totals with NHTS 2017

Data	Total Daily Count
Mobile Device OD	18,007,233
NHTS Longer than 50 Miles	24,608,128
NHTS Longer than 75 Miles	15,016,246
NHTS Longer than 100 Miles	10,526,320

Figure 21 shows the comparison of trip generation between the national weekday person OD products and the NHTS 2017 trips longer than 75 miles. We can see that the two products are consistent. **Figure 22** is also focused on trip generation, but only includes air trips. The comparison is between the national person OD for air trips and DB1B for the first quarter of 2017. The difference here is that the mobile device OD represents actual origins, while DB1B is airport-based. As a result, in states such as Virginia and Illinois whose airports attract travelers from their neighboring states, we see higher shares in DB1B.

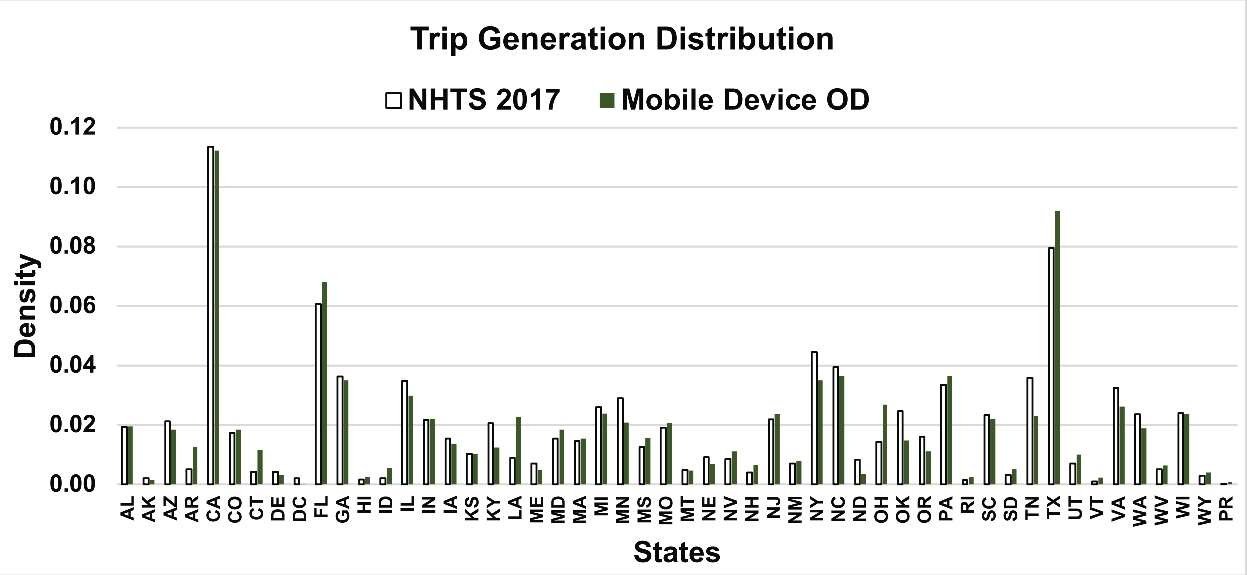


Figure 21. Comparison of trip generation by state between weekday OD and NHTS 2017 for trips longer than 75 miles

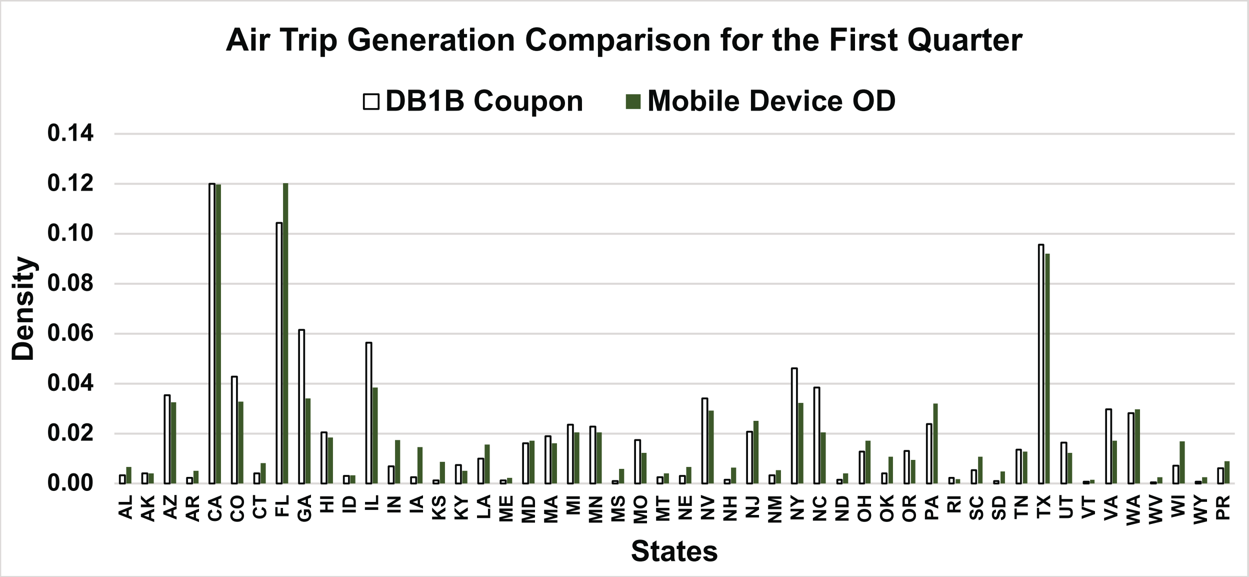


Figure 22. Comparison of first quarter average day OD of air trips with DB1B

Figure 23 shows the comparison of trip distance distribution between the national weekday person OD products and the NHTS 2017 trips longer than 75 miles. A satisfactory match is observed.

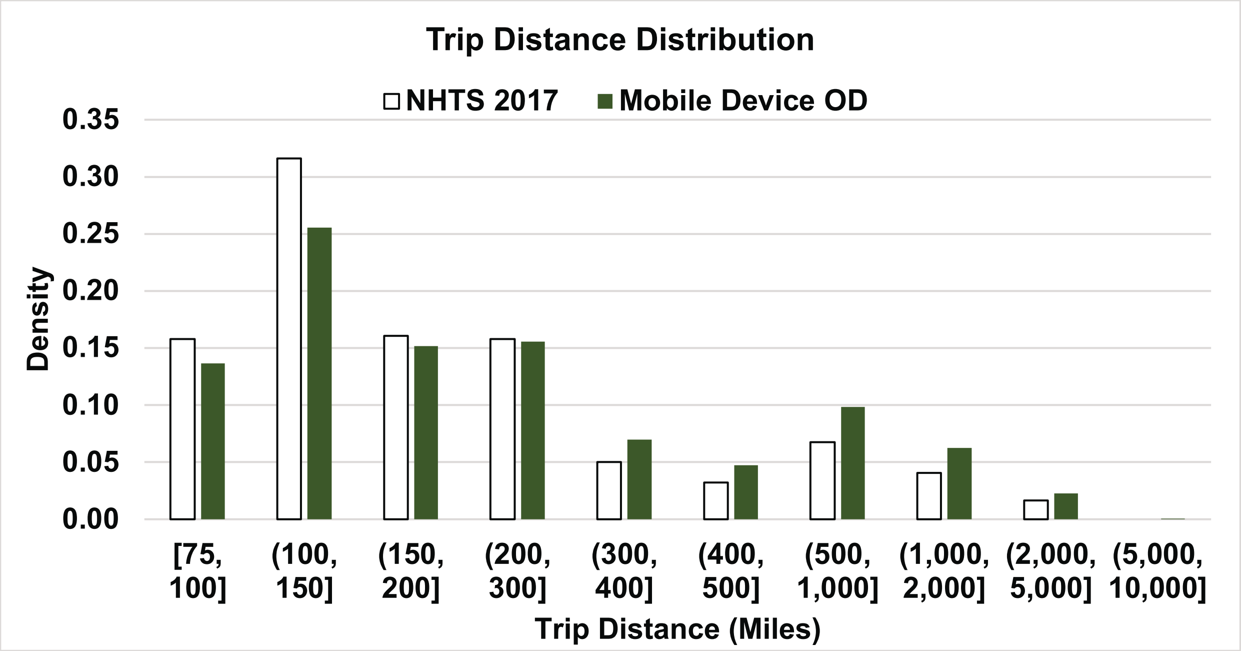
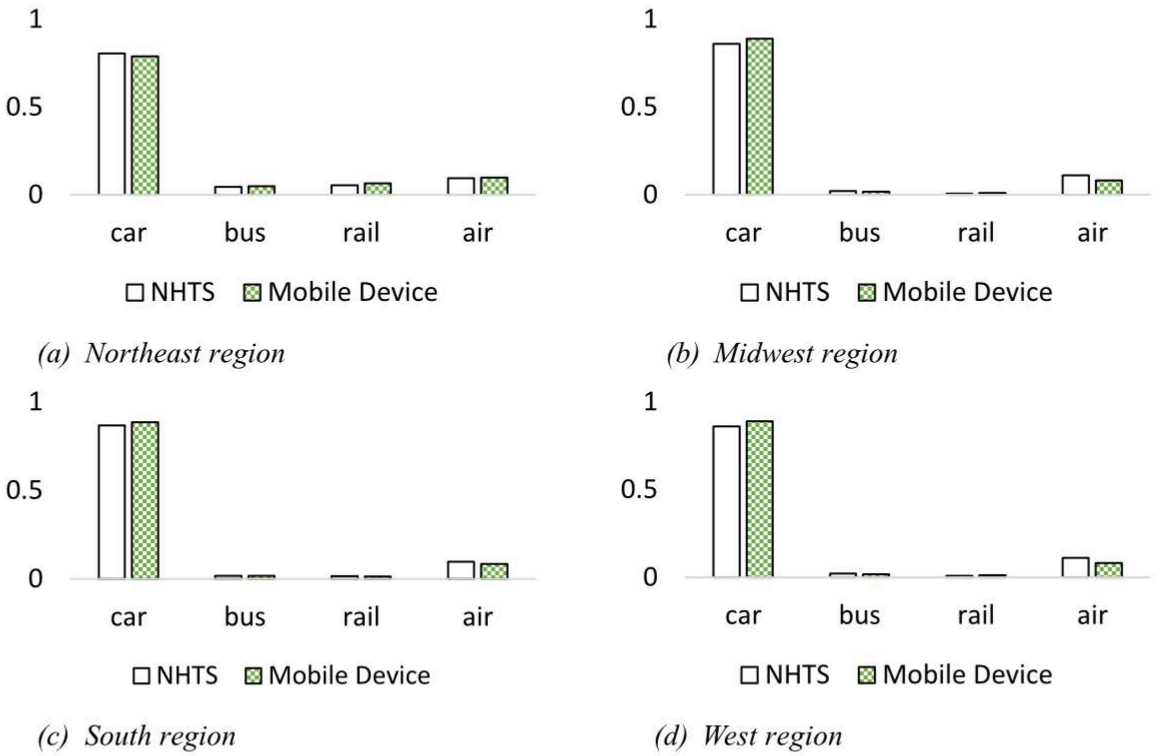


Figure 23. Comparison of trips distance distribution between weekday OD and NHTS 2017 for trips longer than 75 miles

Figure 24 shows the mode share comparison between the national weekday person OD products and the NHTS 2017 trips longer than 75 miles for different census regions. We can see that the two datasets match well over all regions.



(a) Northeast region

(b) Midwest region

(c) South region

(d) West region

Figure 24. Comparison of mode share between weekday OD and NHTS 2017 for trips longer than 75 miles in four different US census regions

Figure 25 shows the comparison of total daily air trips (in millions) between the national person OD, DB1B, and T100. We can see that the three datasets are showing consistent results. We need to note that the OD products are first produced for each month, then aggregated to annual values for the project deliverable.

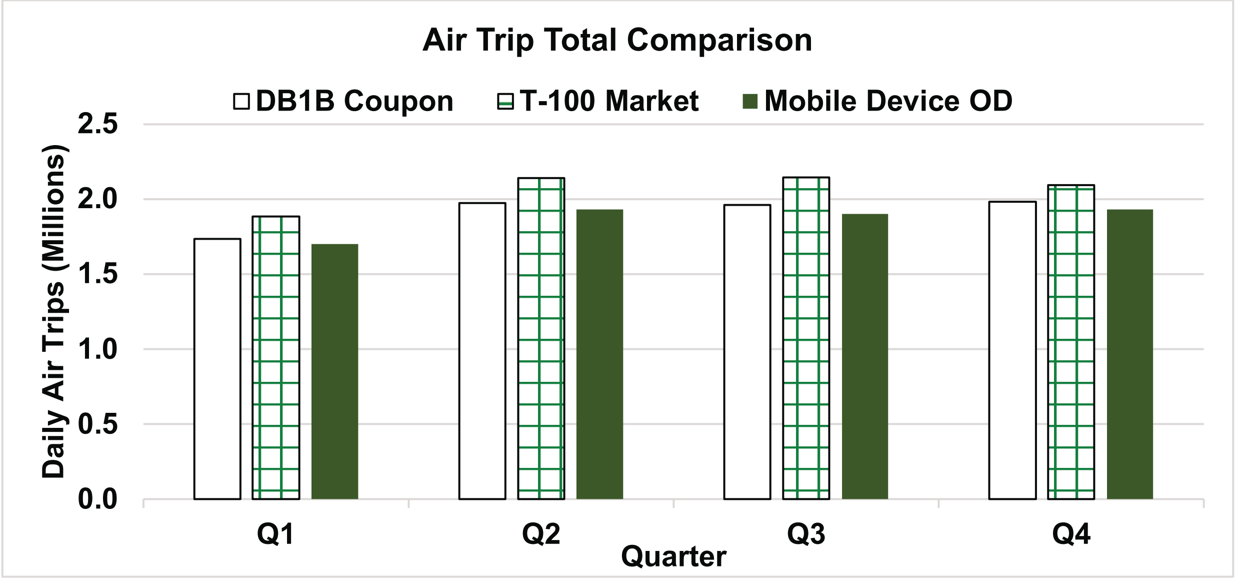


Figure 25. Comparison of total air trips by quarter between average day OD, DB1B, and T100

5.2. National Truck Trip Product

This product contains U.S. national long-distance truck travel OD data for the year 2017. The product includes annual daily average values for truck travel. Each row of a CSV file corresponds to weighted truck trip counts between one specific OD pair. The information is cross-tabulated by (1) two vehicle weight classes, i.e., Medium (medium duty trucks/ vans: ranges from 14001-26000lb) and Heavy (heavy Duty Trucks: >26000 lb), and (2) three day types, i.e., Average_Day (Monday through Sunday), Average_Weekday (Monday through Friday), and Average_Weekend (Saturday and Sunday). **Table 9** shows the variables in the product with their description and variable type. The count information is presented for each combination of vehicle weight class and day type, leading to six count columns.

Table 9. National truck OD product codebook

Name	Label	Type
Origin_Zone_ID	GeoID code of the origin zone based on the national zone system	Numeric
Origin_Zone_Name	The name of the origin zone	Character
Destination_Zone_ID	GeoID code of the destination zone based on the national zone system	Numeric
Destination_Zone_Name	The name of the destination zone	Character
Count	Weighted number trips between each specific OD pair	Numeric

Table 10 shows the results for the average daily truck trips calculated from the national truck OD table. We can see that the average daily trips in different months of year is close to the value derived from FAF, 1.84 million. The share of the intra-zonal trips in the product is 69%. This share is equal to 0.66 for trip tables derived from FAF. We must note that the OD products are first produced for each month, then aggregated to annual values for the project deliverable.

Table 10. Mobile device OD average daily truck trips (in millions) by month

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Total	1.45	1.37	1.79	1.67	1.84	1.89	1.78	1.99	1.86	1.85	1.73	1.67

Figure 26 shows the comparison of truck trip generation between the EAR truck OD and the derived FAF OD. We can see that the EAR OD product shows a decent match with the FAF OD.

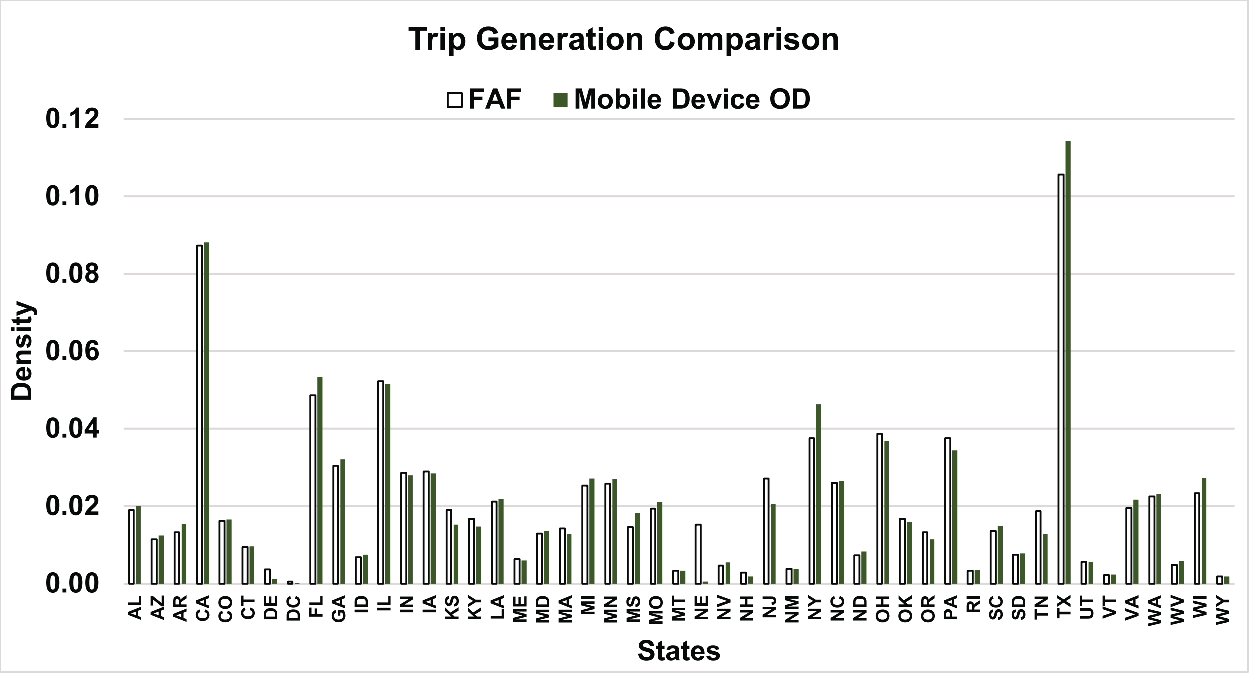


Figure 26. Comparison of trip generation by state between truck OD and FAF

5.3. MPO Person Trip Product

This product contains MPO person travel OD data for the year 2017. The product includes annual daily average values for person travel. Each row of the CSV file corresponds to weighted person trip counts between one specific OD pair. The count numbers are cross tabulated by (1) five day parts, i.e., All_Day (12am-12am), Morning_Peak (6am-10am), Mid_Day (10am-3pm), Afternoon_Peak (3pm-7pm), and Night (7pm-6am), (2) three day types, i.e., Average_Day (Monday through Sunday), Average_Weekday (Monday through Friday), and Average_Weekend (Saturday and Sunday), and (3) four trip purposes, i.e., All_Purpose (all purposes combined), HBW (home-based work), HBO (home-based other), and NHB (non-home-based). **Table 11** shows the variables in the product with their description and variable type. The fourteen count variables in this table are all presented for each combination of trip day type, day part, and purpose, leading to 840 count columns in the dataset.

Table 11. MPO person OD product codebook

Name	Label	Type
Origin_Zone_ID	GeoID code of the origin zone based on the BMC zone system.	Numeric
Origin_Zone_Name	GeoID code of the origin zone based on the BMC zone system.	Character
Destination_Zone_ID	GeoID code of the destination zone based on the BMC zone system.	Numeric
Destination_Zone_Name	GeoID code of the destination zone based on the BMC zone system.	Character
Count	Total trip counts	Numeric
Income_lessthan_20k	Trip counts for travelers that have a household income of less than $20,000	Numeric
Income_20k_50k	Trip counts for travelers that have a household income of $20,000 to $50,000	Numeric
Income_50k_100k	Trip counts for travelers that have a household income of $50,000 to $100,000	Numeric
Income_morethan_100k	Trip counts for travelers that have a household income of more than $100,000	Numeric
Age_lessthan_35	Trip counts for travelers with age less than 35	Numeric
Age_35_65	Trip counts for travelers with age between 35 to 65	Numeric
Age_morethan_65	Trip counts for travelers with age more than 65	Numeric
Male	Trip counts of male travelers	Numeric
Female	Trip counts of female travelers	Numeric
car	Trip counts with car as a primary mode of transportation	Numeric
rail	Trip counts with rail as a primary mode of transportation	Numeric
bus	Trip counts with bus as a primary mode of transportation	Numeric
non-motorized	Non-motorized trip counts	Numeric

Figure 27 shows trip rate and mode share comparison between the weekday MPO person OD products and regional travel surveys. Two regional surveys are used for this comparison; the first is NHTS2017, for which the samples within the Baltimore metropolitan area are filtered and used; the second is BMC's 2007-2008 Household Travel Survey.

Figure 28 shows the comparison of mode share between the weekday MPO person OD products and the BMC's household travel survey, by county. We can see that the results are consistent in all regions.

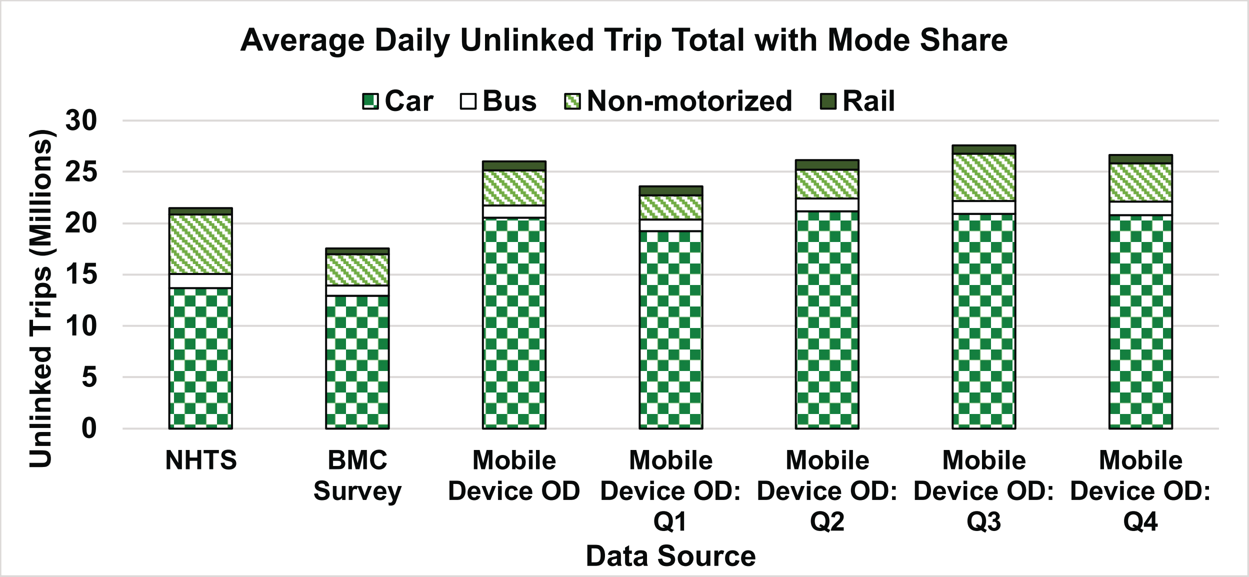


Figure 27. Trip rate and mode share comparison with regional surveys

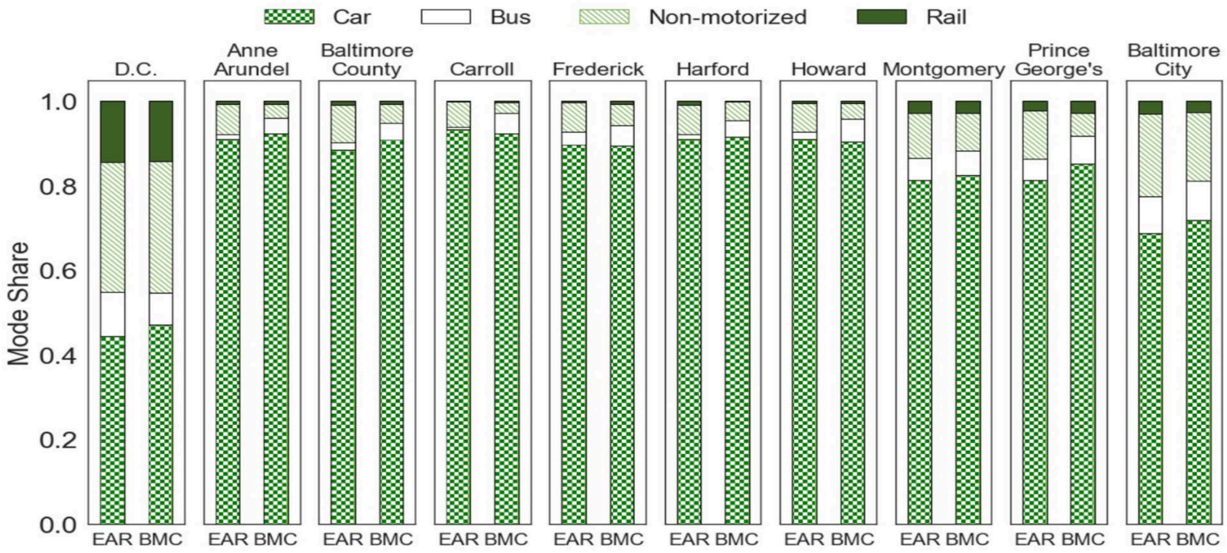


Figure 28. Mode share comparison by county with the BMC survey

Figure 29 shows the trip length distribution between the weekday MPO person OD products and the two regional surveys. We can see that the mobile device OD is underrepresenting trips shorter than one mile. The minimum data points required for identifying a trip is two, if the two data points pass a certain speed, distance, and time criteria. We acknowledge that certain very short trips may not even produce two data points due to the technology limitations; however, the way distances are calculated for the mobile device OD is another contributing factor for observing fewer very short distance trips. The mobile device OD distance distribution is directly calculated from the OD tables using the centroid-to-centroid distances, while the survey distances are from the reported trip distances. The use of centroid-to-centroid distance can contribute to observing fewer very short trips. We should note that the end products are being validated here. The end product, aggregate-level OD tables, do not include trip level information; as a result, only centroid-to-centroid distances available in the end product are used for comparison. We should also note that for the other technology-related contributing factor, future products can apply short-trip rate of high-frequency devices to low-frequency devices in order to address the issue.

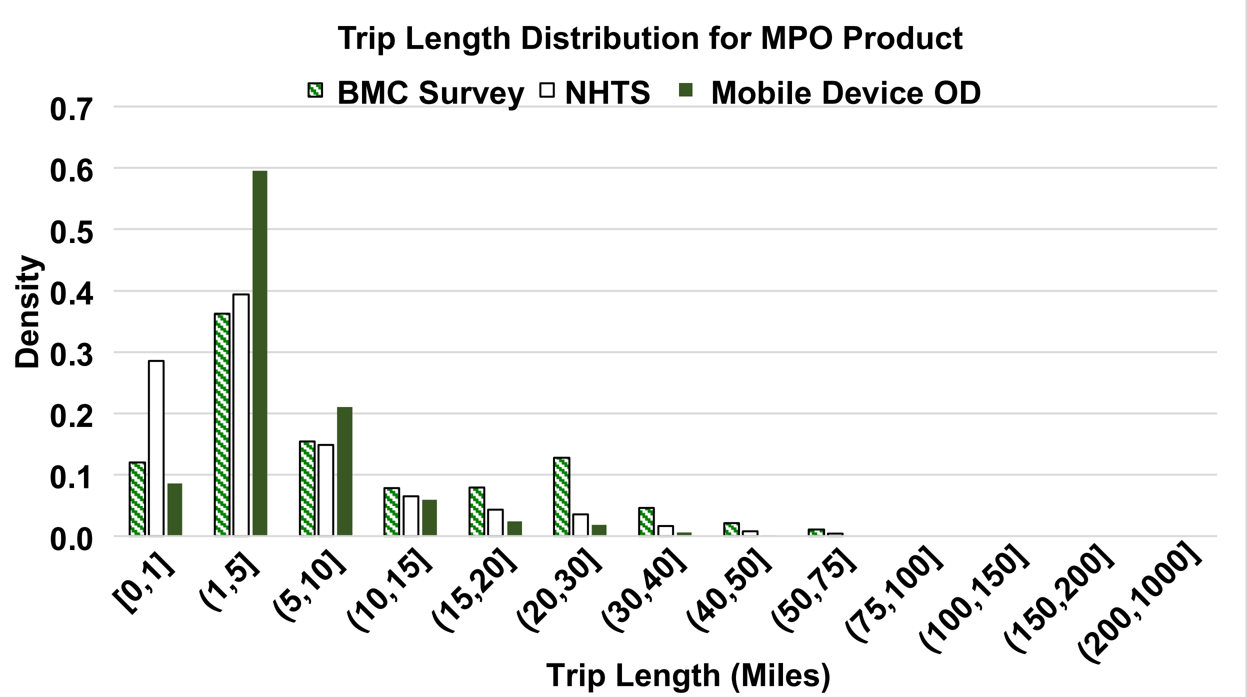


Figure 29. Trip length comparison with regional surveys

Figure 30 shows trip distribution by income comparison between the weekday MPO person OD products and the two regional surveys. A satisfactory match can be seen in the results.

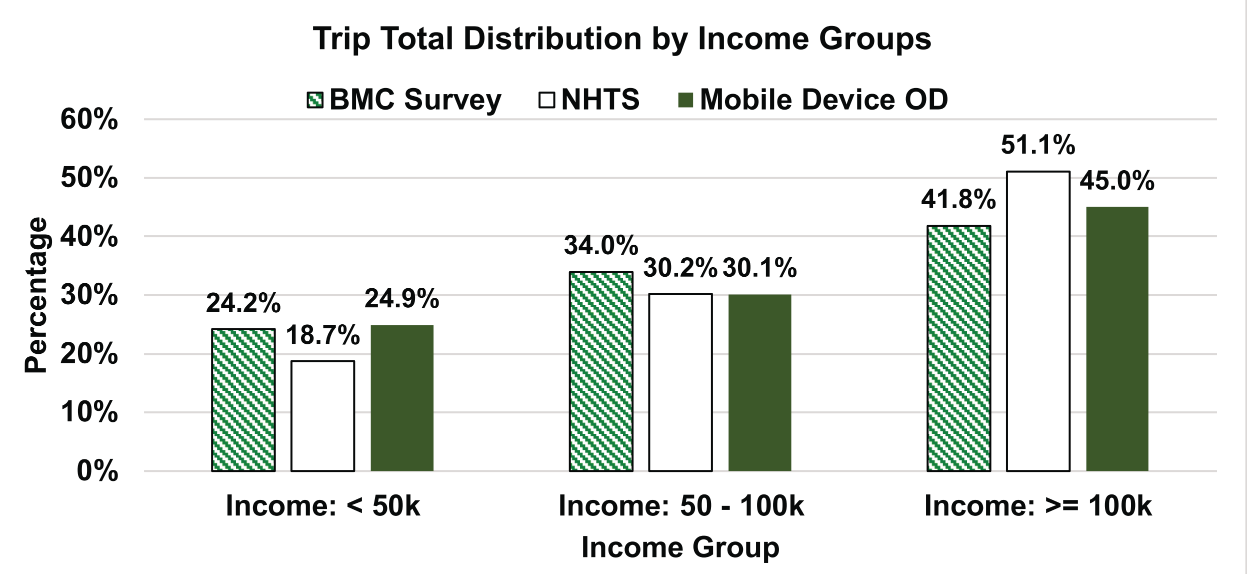


Figure 30. Income distribution comparison with regional surveys

Figure 31 shows trip distribution by age comparison between weekday MPO person OD products and the two regional surveys. A satisfactory match can be seen in the results.

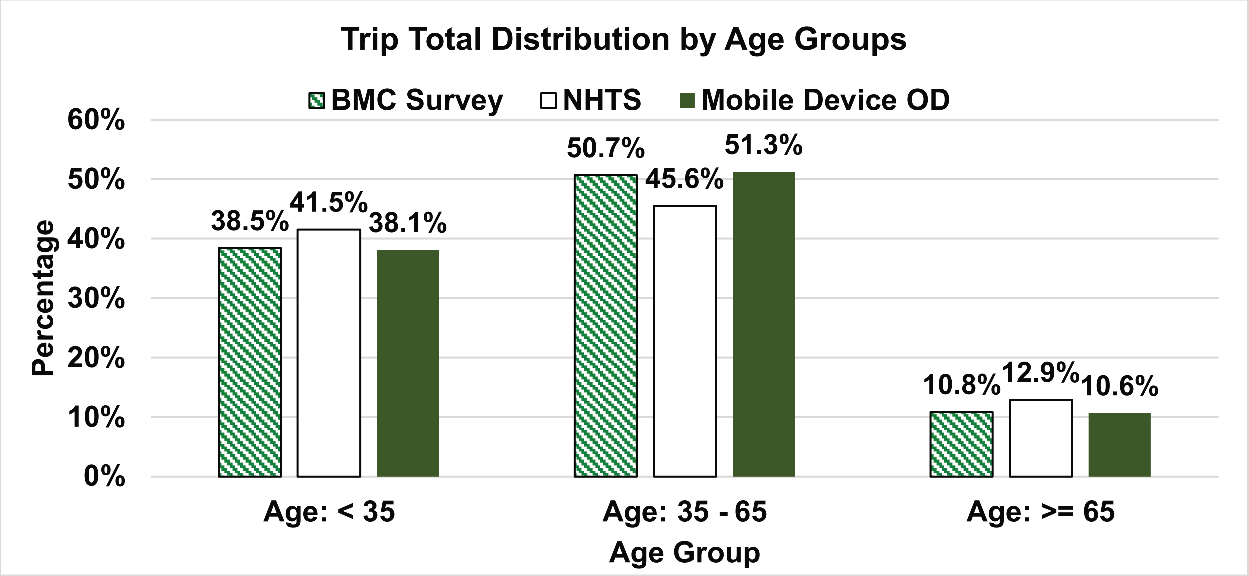


Figure 31. Age distribution comparison with regional surveys

Figure 32 shows transit ridership comparison between the person OD products and the National Transit Database (NTD). NTD only requires data reports from the recipients of FTA (Federal Transit Administration) funds that provide public transportation services. The monthly module of NTD requires ridership data only from the full reporter, defined as urban reporters that either operate more than 30 vehicles for any type of service or that operate 30 vehicles or less on fixed guideway and/or high intensity busway, while our data include many more service providers such as some employer-provided services. As a result, we estimate more bus trips in comparison with NTD. The rail ridership comparison shows satisfactory results.

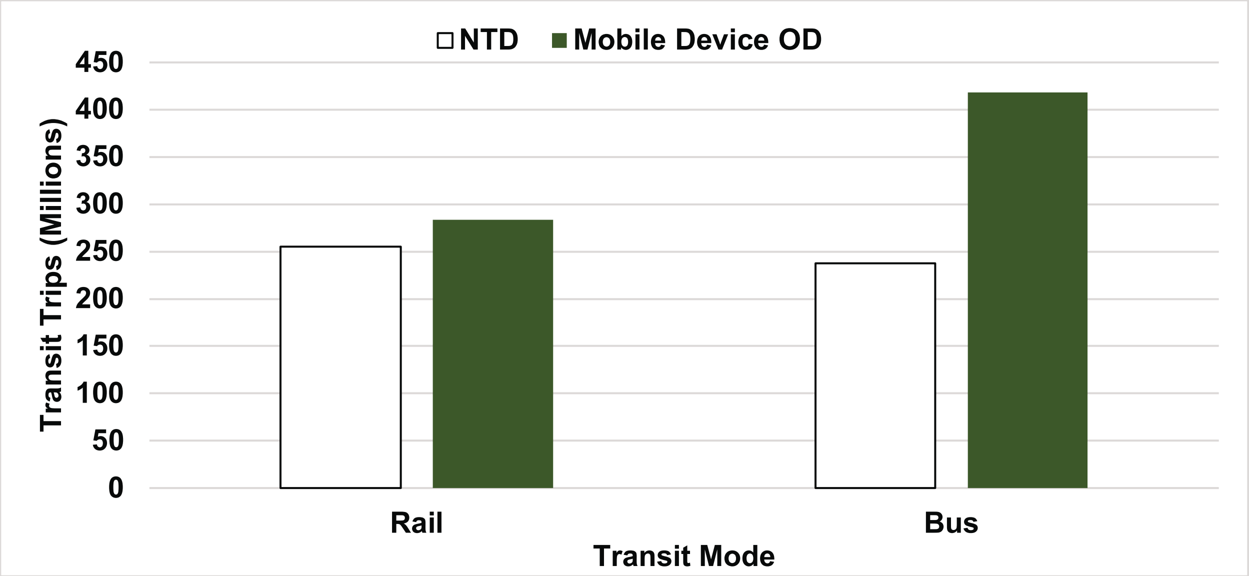


Figure 32. Transit ridership comparison with NTD

5.4. MPO Truck Trip Product

This product contains MPO truck travel OD data for the year 2017. The product includes annual daily average values for truck travel. Each row of the CSV file corresponds to the weighted truck trip count, between one specific OD pair. The count information is cross tabulated by (1) two vehicle weight classes, i.e., Medium (medium duty trucks/ vans: ranges from 14001-26000lb) and Heavy (heavy duty trucks: >26000 lb)), (2) five day parts, i.e., All_Day (12am-12am), Morning_Peak (6am-10am), Mid-Day (10am-3pm), Afternoon_Peak (3pm-7pm), and Night (7pm-6am), and (3) three day types, i.e., Average_Day (Monday through Sunday), Average_Weekday (Monday through Friday), and Average_Weekend (Saturday and Sunday). **Table 12** shows the variables in the product with their description and variable type. The count information is presented for each combination of vehicle weight class, day type, and day part, leading to 30 count columns in the dataset.

Table 12. MPO truck OD product codebook

Name	Label	Type
Origin_Zone_ID	GeoID code of the origin zone based on the BMC zone system	Numeric
Origin_Zone_Name	GeoID code of the origin zone based on the BMC zone system	Character
Destination_Zone_ID	GeoID code of the destination zone based on the BMC zone system	Numeric
Destination_Zone_Name	GeoID code of the destination zone based on the BMC zone system	Character
Count	Weighted number of trips between each specific origin and destination	Numeric

Figure 33 shows the comparison between the MPO truck OD products and the regional freight travel demand model developed as a part of SHRP2 C20 project. The results show a satisfactory comparison.

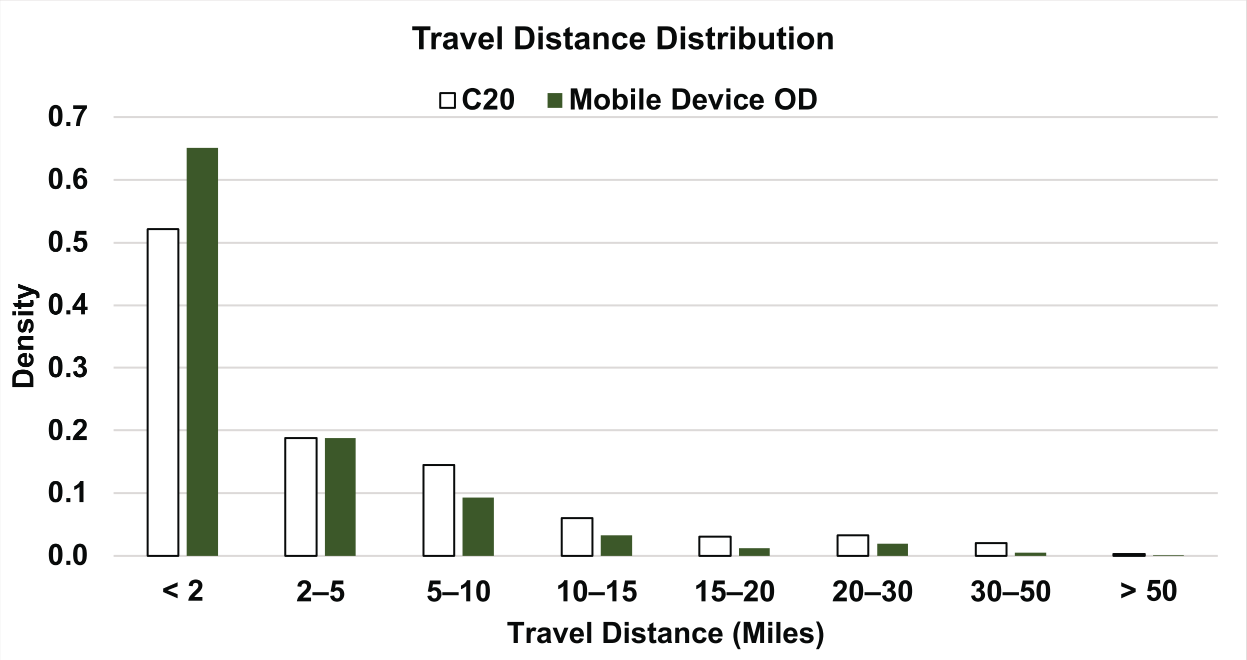


Figure 33. Trip length comparison with Baltimore freight demand model (C20 model)

5.5. Product Differences

In this section, we highlight the differences between the products:

- Intrazonal trips: While the truck OD tables include intrazonal trips, person OD tables are limited to interzonal trips. Including intrazonal trips requires processing all trips, which was not possible in the scope of this project for the person OD tables. Due to significant differences in the input data (in vehicle GPS for truck trips versus LBS for person trips), we were able to process all trips for truck trips within the project scope, therefore, truck OD tables include intrazonal trips.
- Alaska and Hawaii coverage: Person OD tables include Alaska and Hawaii, while truck OD tables are limited to the continental U.S. This is also due to differences of coverage in the input data sources.
- OD pair coverage: The covered OD pairs between the person and truck OD tables are different. Person OD tables cover significantly more OD pairs. As discussed before, the data sources for the two products are innately different, so such differences are inevitable. In this case, the big difference between the covered OD pairs can also be related to the truck trip breaks. While distant OD pairs have observations in the person products, they do not have truck observations.
- Purpose categories: Purpose categories are different between the national person OD tables and MPO person OD tables. National OD tables are limited to long-distance trips; therefore, the purpose categories are set to better represent long-distance trip purposes. Also, purpose is cross tabulated with socio-demographic and mode only in MPO OD products. The cross-tabulation does not exist in the national OD products due to the different methodology used.
- Mode categories: Mode categories between the national and the MPO person OD products are different. National products include air and exclude non-motorized. MPO products exclude air and include non-motorized.

6. ADDITIONAL EXPLORATORY RESEARCH

This FHWA EAR project has benefited from feedbacks of project expert panel members and other external experts through annual expert panel meetings and other review activities. Based on their suggestions, the UMD-led research team has conducted additional exploratory research in support of the project goal, though some of these additional research efforts are out of the scope of the original project work plan. In addition, the research team also voluntarily produced a raw data sandbox to improve data transparency, which is not required by the project work plan. These additional exploratory research efforts are described in this section.

6.1. Truck Trip Chaining

6.1.1. Existing Issues in the GPS Truck Trips

To estimate the truck origin and destination (OD) matrix at both the national and metropolitan planning organization (MPO) level, our research team collaborated with one of the leading data aggregators to obtain large-scale truck GPS data. The raw data include the provider ID, device ID, trip ID, vehicle weight class of each truck, and all the waypoints collected by the in-vehicle GPS devices. In most cases, a new device ID and a new trip ID are generated when the truck engine is turned on, and the trip will end when the engine is turned off. Therefore, the device ID for each truck and the trip ID for each trip may vary and cannot be used to identify the same truck. Moreover, trips with intermediate stops, where the engine is turned off, will be segmented into multiple trip records. Even for the truck GPS data with consistent device IDs, the state-of-the-practice methods only identify trips based on the dwell time at stops and consider those stops with long dwell times as the actual trip end. They usually neglect the intermediate stops at truck parking lots and gas stations, where trucks stop for resting purposes but do not make freight delivery. These issues lead to a biased truck demand estimation, as long-distance and interstate truck trips might be considered as multiple short-distance trips.

6.1.2. Proposed Methods

To address this issue, our team proposed a truck trip chaining algorithm to chain the initial trip records that start from and/or end at truck parking facilities and/or gas stations. According to the Federal Motor Carrier Safety Administration (FMCSA), truck drivers must be allowed 10 hours of off-duty time after completing an 11- to 14-hour on-duty period. Our team has also taken this rule into consideration.

1. Select trips with ends close to truck parking lots (Subset Ⅰ) or gas stations (Subset Ⅱ): 1 mile for truck parking lots and 0.25 mile for gas stations.
2. For each subset, sort trips chronologically by trip start time, and start from the first trip (the target trip).
    a. Find all candidate trips with the same provider ID and vehicle weight class.

   b. If there are candidates with the same device ID as the target trip, calculate the time difference between the end time of the target trip and the start time of those candidates with the same device ID.
   c. If there is no candidate trip with the same device ID, calculate the time difference between the end time of the target trip and the start time of all the candidate trips that depart later than the end of the target (previous) trip.
   d. In addition, our team evaluates the FMCSA rule for trips with ends close to truck parking lots: if the driving time of the target trip is longer than 11 hours, only consider the candidate trips that depart at least 10 hours later than the end of the target (previous) trip.
   e. Link the target trip with the closest candidate trip and remove those two trips from the subset.
   f. Move to the next unmatched trip.
3. After the primary trip linking for Subset Ⅰ and Ⅱ, further link the trip chains by trip ID so that a trip with multiple stops at truck parking lots and/or gas stations would be recovered.

The aforementioned method uses the temporal difference as the major measurement to link truck trips in addition to the intuitive device ID, provider ID, and vehicle weight class. The research team also considers two GPS data features, the tracking frequency and the travel speed profile, that are potentially helpful. The motivation is twofold: 1) the tracking frequency of the same GPS device should be consistent, and 2) the driving speed of the same truck driver should be consistent. For the addition of the GPS data features, some possible challenges are to construct a similarity index that measures the similarity of tracking frequency and speed between trips, and to determine the similarity threshold to link truck trips. In our sample data, nearly 80% of truck trips share the same device IDs with at least one other trip. Therefore, we would be able to utilize them as a training dataset and find the optimal thresholds by trial and error.

6.1.3. Before-and-After Comparison

In this section, our team will deliver some preliminary results to demonstrate the feasibility and efficiency of the proposed method based on the temporal difference (excluding the GPS data features). Our team applied the trip chaining algorithm to one month (June 2015) of truck trip data that have waypoints observed in the state of Maryland. After trip linking, the number of OD pairs with observed trips increases from 2,864 to 3,438, indicating that the truck trip chaining algorithm recovers more OD pairs, especially with long distance. **Figure 34** summarizes the distribution of trip distances between origin and destination zone centroids. It can be seen that the number of trips less than 100 miles is significantly reduced and more long-distance trips are observed.

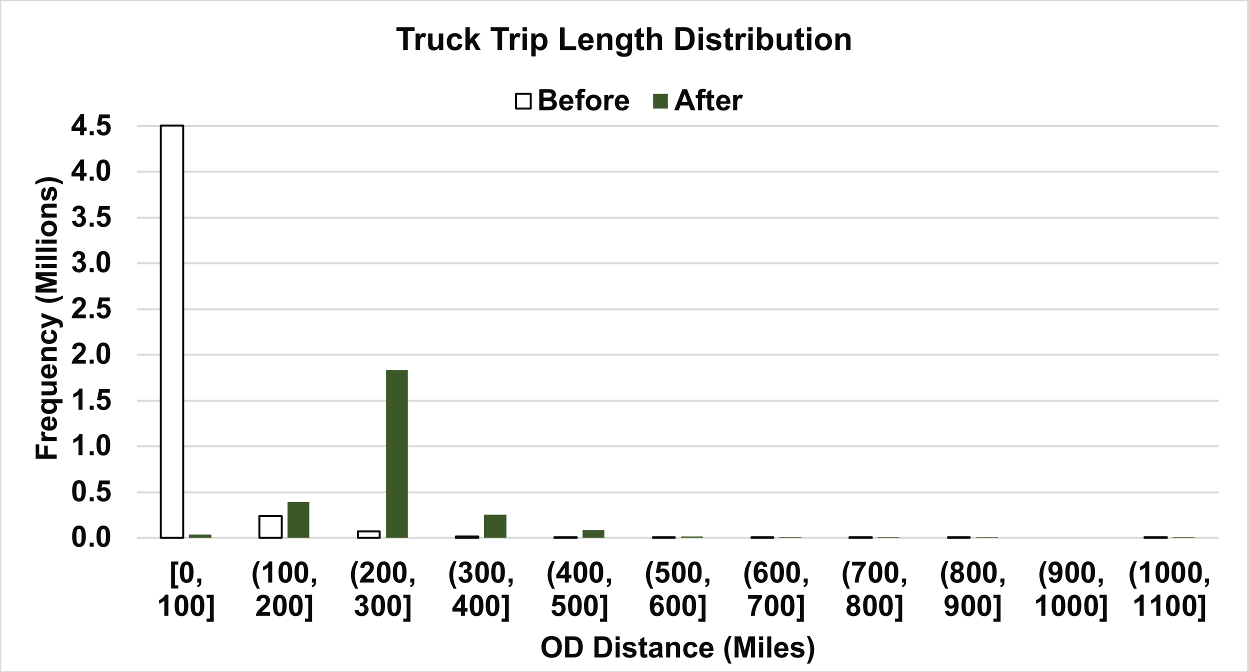


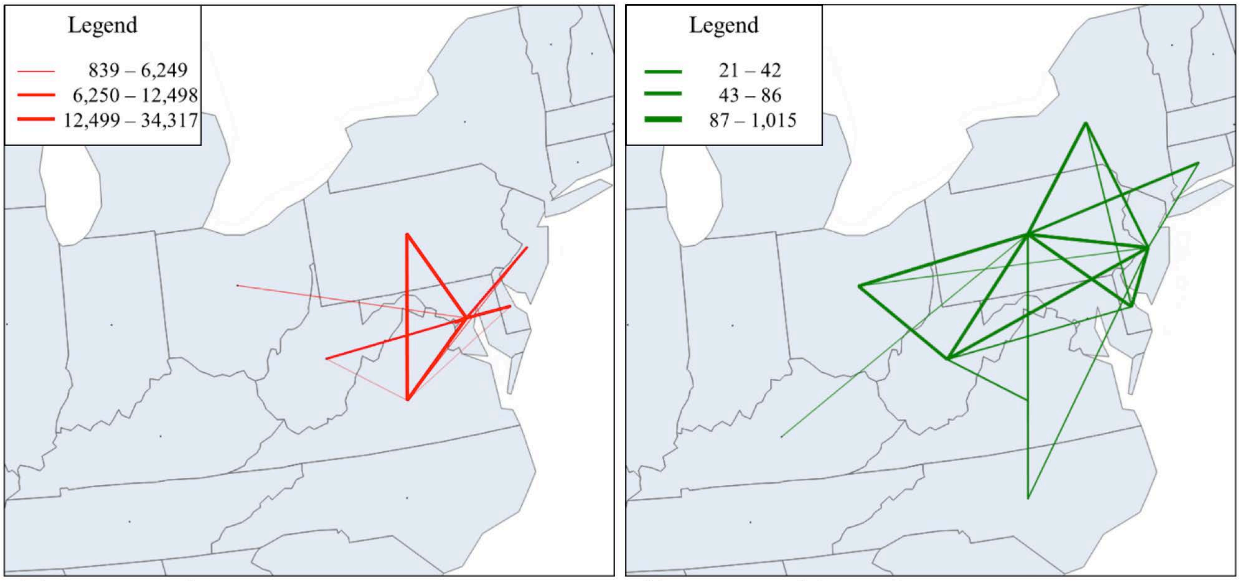
Figure 34. Trip distance distribution before and after truck trip chaining

Table 13 compares the ratio of intrazonal trips in the state of Maryland before and after applying the truck trip chaining algorithm. The benchmark estimate is based on the converted truck trips from the 2017 Freight Analysis Framework (FAF) tonnage data. The truck trip chaining algorithm has significantly decreased the intrazonal trip ratio from 75% to 69% by linking truck trips together; the after-linking result is much closer to the estimate based on FAF.

Table 13. The ratio of intrazonal trips before and after truck trip chaining

	MD Trip Estimates		Intrazonal Trip Ratio Estimates		
Scenario	**Intrazonal trips**	**All trips**	**MD estimate based on mobile device data**	**MD estimate based on FAF**	**National estimate based on FAF**
Before	3,632,465	4,835,882	75%	54%	66%
After	2,962,912	4,270,927	69%		

We also compared the state OD differences before and after applying the truck trip chaining algorithm. **Figure 35** (a) and **Figure 35** (b) displayed the state OD pairs with the top 25% decrease and increase in the number of trips, respectively. For instance, the number of truck trips between Maryland and neighboring states has reduced, and the number of long-distance truck trips between Pennsylvania, North Carolina, and Tennessee has increased. The results indicate that truck trips with both ends outside Maryland were underestimated before trip chaining.



(a) Decreased State OD Pairs *(b) Increased State OD Pairs*

Figure 35. State OD pairs with top 25% trip changes

Overall, the after-linking results imply a better demand estimation compared with the external data sources and the before-linking results.

6.2. Socio-Demographic Imputation

This section describes additional research efforts on socio-demographic imputation based on the overall method described in **Figure 18**.

6.2.1. Feature Engineering

The focus of feature engineering is two-fold: spatial temporal dynamics and contextual semantics of locations. In addition to the simple statistics of relevant features, more advanced representations are also reviewed and examined, such as an embedded representation of stay points (Solomon, et al., 2018) and spatial temporal entropy sequence (Moro, et al., 2018). In feature set construction, four types of information are considered: the travel behavior characteristics, geographic information for imputed home and work locations, POI information, and imputed trip purposes.

6.2.2. Classification Methods

Previous studies have tested and compared most of the prevalent classification methods, such as linear discriminant analysis (LDA), logistic regression, support vector machine (SVM), partial decision tree classification (PART), random forest, Bayesian inference method, generalized additive model, etc. The model performances vary between different datasets. In addition to the single classification model, other complex frameworks were also proposed including multi-task learning (Wang, et al., 2016; Zhong, et al., 2013) and multi-level classification model (Ying, et al., 2012). Since the target is to impute multiple attributes for the same device, the structured learning methods were also evaluated. To address the classification problem, the conditional inference tree

(CIT) model and CIT-based random forest are tested because of their abilities to avoid biased variable selection and overfitting. In addition, there exist sample imbalance problems across different classes, so the examination includes parallel models with weight adjustment.

6.2.3. Results

To evaluate the model performance, a seven-fold cross-validation is employed. To make the results comparable, the random seeds to generate the bootstrap samples are fixed. In addition to the overall accuracy of the prediction, other measurements including recall, precision, and F1 score are also calculated. Recall is the fraction of relevant instances that have been retrieved over the total amount of relevant instances (i.e., the proportion of correctly imputed instances) and precision is the fraction of relevant instances among the retrieved instances. F1 score is the harmonic mean of precision and recall ($F_1 = 2 \times \frac{precision \times recall}{precision + recall}$), the best of which is one. The overall accuracy is the proportion of correctly imputed instances for all the groups.

Table 14 displays the best model for each attribute concerning the overall accuracy with the F1 score. It can be observed that the overall accuracy is better than random assignment but indicates plenty of room for improvement. For example, the F1 scores imply a model bias to the class with more instances (35+ and medium income group).

Table 14. Socio-demographic imputation results

Attribute	Feature Set	Method	F1 Score			Overall Accuracy
Age Group	Travel behavior and geographic information	Weighted conditional inference tree (CIT)	Under 35	35+		74.62%
			35.83%	84.59%		
Income Level	All four types of information: travel behavior, geographic information, POI information, and imputed purposes	Random Forests	Low	Medium	High	53.44%
			12.90%	67.54%	29.36%	

This is an ongoing process. We are trying to add new features and fine-tuned the models to get a higher accuracy and improve the imputation results.

6.3. Raw Data Sandbox

6.3.1. What Is the Raw Data Sandbox?

The research team produced a sample of raw location data from mobile devices, called the Raw Data Sandbox. The Raw Data Sandbox was prepared to allow various data users and the broader researcher and practitioner community to better understand raw mobile device location data and improve their confidence and appreciation of relevant data products (e.g., origin-destination tables and others). The Raw Data Sandbox is a sample of raw data, anonymized and aggregated to small zones to protect privacy. We must note that all raw data flaws, such as duplicate observations and location jumps, are intentionally left as-is in the Raw Data Sandbox, so that the potential users can also have a better understanding over data issues and limitations.

6.3.2. What Can Users Do with the Raw Data Sandbox?

The sandbox includes samples of raw location data for people and vehicles from different data providers. These samples benefit the users by showing them the raw data structure, data frequency, data coverage, and data flaws. The samples can be used to develop and test algorithms for data cleaning, trip-identification, sample weighting, and mode/purpose/socio-demographic imputations. The users can also submit algorithms developed using the data sandbox or tested on the data sandbox to be applied on the protected full sample and get the aggregate-level validation back. This process can help engage the entire research community in advancing the methodologies for utilizing mobile device location data.

6.3.3. What Data Are Included in the Raw Data Sandbox?

The Raw Data Sandbox includes anonymized raw location data for both person location and vehicle trips. The geographical coverage of the data in the Raw Data Sandbox is the Baltimore metropolitan area.

- Person location data: Person location data are generated by the location-based services (LBS) within mobile devices. LBS data are obtained from the interaction between smartphone apps and software development kits that are designed to record the device location. The data sandbox includes data from multiple data providers. An undisclosed share/subsample of all devices from each original data provider is included in the data sandbox, which also protects business-sensitive information for data providers. The temporal coverage is one week in July 2017 (July 23~29, 2017).

- Vehicle trip location data: Vehicle trip location data are generated from the GPS devices inside vehicles. The data include both passenger vehicles and trucks. The GPS devices frequently record the location of the vehicle and produce sightings. GPS location data from one major GPS data provider, separated by passenger cars and trucks, are included in the sandbox. The temporal coverage is one week in July 2018 (July 23~29, 2018).

The data provider partners contributing to the Raw Data Sandbox are AirSage, INRIX, StreetLight Data, and Cuebiq.

6.3.4. How Does the Raw Data Sandbox Protect User Privacy?

After an extensive discussion between MTI, FHWA, and data provider partners, the team decided to aggregate the raw location data to small hexagons to address privacy concerns. The original location data were aggregated to hexagons that cover the entire world based on the Uber H3 indexing system (Uber's introduction to H3). Each location point was substituted with a hexagon index. Each hexagon index represents a hexagon covering a defined area on the map. The size of the hexagons depends on the H3 resolution, which can lead to small or large hexagons. In the Raw

Data Sandbox, H3's resolution 7 was used, which divided the Baltimore metropolitan area into around 6500 zones. At this resolution level, the length of each edge of a hexagon zone is 1.22 km. The hexagons for the Baltimore metropolitan area can be seen in **Figure 36**.

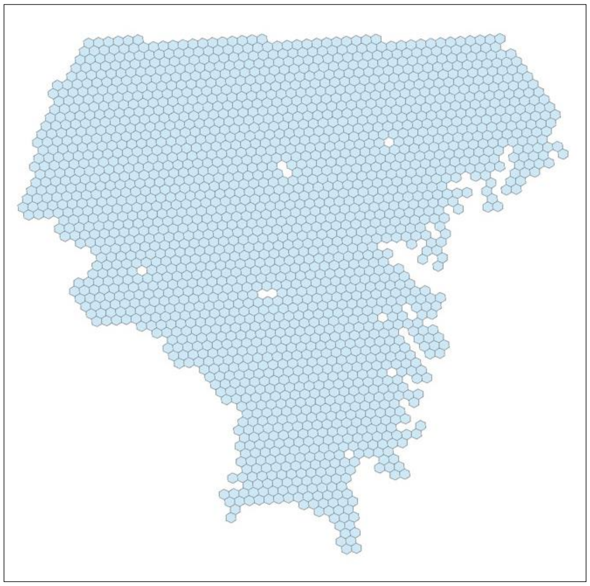


Figure 36. H3 resolution 7 hexagon zones for the Baltimore metropolitan area

After this aggregation process, if the original dataset includes fewer than 10 devices inside a hexagon, all sightings inside the hexagon were removed from the Raw Data Sandbox to further protect privacy.

6.3.5. What does the Raw Data Sandbox Look Like?

The Raw Data Sandbox has two components: the person location sandbox and the vehicle trip sandbox. The structure of each component is described below:

- *Person Location Sandbox:* This component is a CSV file containing anonymized location data from mobile devices. Each row of the file represents one observation of one mobile device. Hashed device ID, time stamp, and hexagon ID are available for each location data observation. Hexagon IDs are based on Uber H3 zone system (Resolution 7). **Table 15** shows a sample row of the person location sandbox.

Table 15. Cellphone data sandbox format

Device_ID	Time_stamp	Hexagon_ID
46f046aaceca8fec2770ce	2017-07-24 12:35:06	87f042129ffffff

- *Vehicle Trip Sandbox:* The raw location data from in-vehicle GPS devices form this sandbox component. The in-vehicle GPS data are different from the person location data in that these locations come in a trip trajectory format. For each trip, the origin, the destination, and the waypoints are available in the original dataset. The vehicle trip sandbox itself has two components: passenger cars and trucks. For each component, the sandbox includes a trip CSV file which has information on trip origin (aggregated to Uber H3 resolution 7 hexagons), trip destination (aggregated to Uber H3 resolution 7 hexagons), hashed device ID, hashed trip ID, and time stamp. Each component's sandbox also includes a waypoint CSV file that has information on waypoints (latitude, longitude, timestamp, trip ID, device ID). The waypoints that fall into the origin and destination hexagons are removed from the sandbox to protect privacy. The waypoint file can be linked to the trip file using the trip ID. **Table 16** and **Table 17** show sample rows of the trip CSV file and the waypoint CSV file, respectively.

Table 16. GPS data sandbox trip file format

Trip_ID	Device_ID	Start_Time	End_Time	Origin_Hexagon	Destination_Hexagon
22cf4ff57	0763a7	2018-07-23T16:30:07	2018-07-23T17:04:07	87f04216f	87f042a93ffffff

Table 17. GPS data sandbox waypoint file format

Trip_ID	Device_ID	Time	Latitude	Longitude
22cf4ff57	0763a7	2018-07-23T16:38:07	39.27155	-76.7228

The data sandbox directory includes two folders, one for person locations and one for vehicle trips. The vehicle trips folder is further divided into two folders, one for passenger cars and one for truck. Each folder includes the required sandbox CSV files in addition to a readme file. The shapefile for the study area and the Uber H3 hexagons is also available in the sandbox directory.

7. CONCLUSIONS

The UMD team, including data scientists, travel behavior analysts, algorithm developers, survey experts, providers of passively collected data, and state department of transportation (DOT) and metropolitan planning organization (MPO) partners, has explored and confirmed the feasibility of producing high-quality, person and truck travel OD tables with transparency and user confidence in both the source data and computational methods. We have demonstrated the feasibility of producing quality person and truck travel OD tables at the national and statewide/MPO levels based on available mobile device location data. Computational algorithms have been developed and implemented to address known issues with mobile device data such as sample bias and missing trip and traveler information. Validation results show that OD tables based on mobile device data are consistent with control totals derived from established independent data sources.

The project contributes to the field by producing accurate products and algorithms supported by robust validation. The project also contributes to improved transparency and openness by producing the Raw Data Sandbox and open-source computational algorithms for data processing, trip identification, imputation, weighting, and validation. Future research may focus on establishing national standards or guidelines on raw data quality, data and method transparency, computational algorithm performance, and validation targets. FHWA and other agencies may consider integrating OD products from mobile device data into their existing data programs and business processes.

The team faced several challenges in this project, which were overcome through innovation, data provider support, and agency guidance. The main challenges and our solutions are summarized below, which we hope would inspire future research and further enhanced solutions.

1) Balancing privacy and transparency: Location data that are continuously being collected from mobile devices may reveal sensitive personal information. Users are legitimately concerned about their privacy. Agencies are acknowledging the concerns and responding to them by putting more restrictions on passive data collection. As researchers, we are dedicated to protecting users' privacy. On the data provider side, businesses are legitimately protecting their businesses and trying to be competitive through their proprietary methods and procedures. This has led to a lack of confidence from the users, who see the system producing the products as a black box. As a research team working with passively collected mobile device location data, we try to promote transparency as much as possible. Balancing transparency and privacy protection can be a challenge. The Raw Data Sandbox can serve as a great example of balancing privacy and transparency.

2) Dealing with data limitations and data quality issues: Surveys are designed to include the required information for answering questions of interest. On the other hand, passively collected location data lack any direct information about the subject or the context. As a result, they have limited details to answer questions of interest. Furthermore, the collected data are bound by technology limitations. Duplicate observations, location jumps, and inaccurate observations are still part of the data. The collected data vary between regions, time-periods, and devices. Methods developed on higher quality observations may not work with the lower quality observations and the methods developed for the lower quality

observations may not fully exploit the potential of the higher quality observations. Over the course of this project, we developed a better understanding of the data limitations and tried to address them through various algorithms. We believe the value of transparent raw data quality metrics and the publication of these metrics for any raw dataset used in the production of OD tables.

3) Dealing with unstructured and unlabeled data: Despite many novel studies on applications of machine learning in transportation, imputation of unobserved information from mobile device location data is still a new field. Advanced machine learning methods that work well on images or texts may not necessarily work well with location data. Besides the methodological challenges, a lack of training data also makes it hard to develop imputation algorithms. The researchers have been utilizing unsupervised learning methods or using limited labeled data that are collected in small scales at specific regions (e.g., smartphone-based travel surveys), but the field can certainly benefit from more ground truth labeled data.

4) Addressing sample biases: Even though the biggest benefit of mobile device location data is their sample size, they still do not cover the entire population. The collected data is a sample, and in many cases, not a representative sample. There are known biases toward certain population groups or certain geographic regions. In addition to the fact that the collected data only represent a sample of the population, the collected data fail to give a complete picture of the covered devices' travel pattern. Only a non-representative sample of all trips is recorded for each observed device. Dealing with sample extrapolation and weighting is a challenging task, requiring data about ground truth control totals, which may not exist with the desired details in many cases. Our approach recognizes both device-level biases and trip-level biases. A multi-level weighting method should be pursued.

5) Comprehensive validation: There are various ground truth data about different aspects of the OD tables, such as trip rate, mode share, trip length distribution, etc.; however, no single dataset exists for a comprehensive and consistent validation. Different datasets may have inconsistent definitions, methods, or assumptions, which can make a comprehensive validation challenging. Converting linked transit trips to unlinked transit trips, converting vehicle trips to person trips, or dealing with access/egress are the sort of tasks that must be taken care of to make different datasets consistent and comparable with the products from the mobile device location data. While a number of validation tests were performed in this project, agencies may consider establishing a standardized validation target to improve transparency and user confidence.

6) Balancing accuracy and computation burden: More advanced algorithms may increase accuracy, but the accuracy gain may not be always worth the extra computation burden and computation costs. Evaluating if the accuracy gain is worth the extra computation burden can be a challenging decision that may depend on the application, computational resources, and the overall goals. This further highlights the importance of agreed-upon product quality standards, which would provide guidance on investment in raw data and computation.

REFERENCES

Altshuler, Y., N. Aharony, M. Fire, Y. Elovici and A. Pentland. (2012), 'Incremental Learning with Accuracy Prediction of Social and Individual Properties from Mobile-Phone Data', (Ed.)^(Eds.), *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, IEEE.

Amin, S., S. Andrews, S. Apte, J. Arnold, J. Ban, M. Benko, R. M. Bayen, B. Chiou, C. Claudel and C. Claudel. (2008), 'Mobile Century Using Gps Mobile Phones as Traffic Sensors: A Field Experiment'.

Antoniou, C., C. L. Azevedo, L. Lu, F. Pereira and M. Ben-Akiva. (2015), 'W–Spsa in Practice: Approximation of Weight Matrices and Calibration of Traffic Simulation Models', *Transportation Research Procedia* Vol. 7, pp. 233-253.

Auld, J., A. Mohammadian, M. Simas Oliveira, J. Wolf and W. Bachman. (2015), 'Demographic Characterization of Anonymous Trace Travel Data', *Transportation Research Record: Journal of the Transportation Research Board*, No. 2526, pp. 19-28.

Balakrishna, P., R. Ganesan and L. Sherry. (2008), 'Airport Taxi-out Prediction Using Approximate Dynamic Programming: Intelligence-Based Paradigm', *Transportation Research Record: Journal of the Transportation Research Board*, No. 2052, pp. 54-61.

Balakrishna, R., M. Ben-Akiva and H. Koutsopoulos. (2007), 'Offline Calibration of Dynamic Traffic Assignment: Simultaneous Demand-and-Supply Estimation', *Transportation Research Record: Journal of the Transportation Research Board*, No. 2003, pp. 50-58.

Bao, J., D. Lian, F. Zhang and N. J. Yuan. (2016), 'Geo-Social Media Data Analytic for User Modeling and Location-Based Services', *SIGSPATIAL Special* Vol. 7, No. 3, pp. 11-18.

Bar-Gera, H. (2007), 'Evaluation of a Cellular Phone-Based System for Measurements of Traffic Speeds and Travel Times: A Case Study from Israel', *Transportation Research Part C: Emerging Technologies* Vol. 15, No. 6, pp. 380-391.

Batini, C., C. Cappiello, C. Francalanci and A. Maurino. (2009), 'Methodologies for Data Quality Assessment and Improvement', *ACM computing surveys (CSUR)* Vol. 41, No. 3, pp. 1-52.

Bierlaire, M., J. Chen and J. Newman. (2013), 'A Probabilistic Map Matching Method for Smartphone Gps Data', *Transportation Research Part C: Emerging Technologies* Vol. 26, pp. 78-98.

Bohte, W. and K. Maat. (2009), 'Deriving and Validating Trip Purposes and Travel Modes for Multi-Day Gps-Based Travel Surveys: A Large-Scale Application in the Netherlands', *Transportation Research Part C: Emerging Technologies* Vol. 17, No. 3, pp. 285-297.

Bohte, W., K. Maat and B. van Wee. (2007), 'Residential Self-Selection. The Effect of Travel-Related Attitudes and Lifestyle Orientation on Residential Location Choice; Evidence from the Netherlands', (Ed.)^(Eds.), *11th World Conference on Transport ResearchWorld Conference on Transport Research Society*.

Boyd, J. H. and R. E. Mellman. (1980), 'The Effect of Fuel Economy Standards on the Us Automotive Market: An Hedonic Demand Analysis', *Transportation Research Part A: General* Vol. 14, No. 5-6, pp. 367-378.

Byon, Y.-J., B. Abdulhai and A. S. Shalaby. (2007), 'Impact of Sampling Rate of Gps-Enabled Cell Phones on Mode Detection and Gis Map Matching Performance', (Ed.)^(Eds.).

Caceres, N., L. M. Romero, F. G. Benitez and J. M. del Castillo. (2012), 'Traffic Flow Estimation Models Using Cellular Phone Data', *IEEE Transactions on Intelligent Transportation Systems* Vol. 13, No. 3, pp. 1430-1441.

Calabrese, F., M. Colonna, P. Lovisolo, D. Parata and C. Ratti. (2011), 'Real-Time Urban Monitoring Using Cell Phones: A Case Study in Rome', *IEEE Transactions on Intelligent Transportation Systems* Vol. 12, No. 1, pp. 141-151.

Chen, C., H. Gong, C. Lawson and E. Bialostozky. (2010), 'Evaluating the Feasibility of a Passive Travel Survey Collection in a Complex Urban Environment: Lessons Learned from the New York City Case Study', *Transportation Research Part A: Policy and Practice* Vol. 44, No. 10, pp. 830-840.

Chung, E.-H. and A. Shalaby. (2005), 'A Trip Reconstruction Tool for Gps-Based Personal Travel Surveys', *Transportation Planning and Technology* Vol. 28, No. 5, pp. 381-401.

Cortes, C. and V. Vapnik. (1995), 'Support-Vector Networks', *Machine learning* Vol. 20, No. 3, pp. 273-297.

Cottrill, C., F. Pereira, F. Zhao, I. Dias, H. Lim, M. Ben-Akiva and P. Zegras. (2013), 'Future Mobility Survey: Experience in Developing a Smartphone-Based Travel Survey in Singapore', *Transportation Research Record: Journal of the Transportation Research Board*, No. 2354, pp. 59-67.

De Choudhury, M., M. Feldman, S. Amer-Yahia, N. Golbandi, R. Lempel and C. Yu. (2010), 'Automatic Construction of Travel Itineraries Using Social Breadcrumbs', (Ed.)^(Eds.), *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, ACM.

De Montjoye, Y.-A., C. A. Hidalgo, M. Verleysen and V. D. Blondel. (2013), 'Unique in the Crowd: The Privacy Bounds of Human Mobility', *Scientific reports* Vol. 3, pp. 1376.

Deng, Z. and M. Ji. (2010), 'Deriving Rules for Trip Purpose Identification from Gps Travel Survey Data and Land Use Data: A Machine Learning Approach', *Traffic and Transportation Studies 2010*.

Doherty, S. T., N. Noël, M. L. Gosselin, C. Sirois and M. Ueno. (2001), 'Moving Beyond Observed Outcomes: Integrating Global Positioning Systems and Interactive Computer-Based Travel Behavior Surveys', (Ed.)^(Eds.).

Doyle, J., P. Hung, D. Kelly, S. F. McLoone and R. Farrell. (2011), 'Utilising Mobile Phone Billing Records for Travel Mode Discovery'.

Draijer, G., N. Kalfs and J. Perdok. (2000), 'Gps as a Data Collection Method for Travel Research: The Use of Gps for Data Collection for All Modes of Travel', (Ed.)^(Eds.), *Transportation Research Board 79th Annual Meeting*.

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

Elango, V. and R. Guensler. (2010), 'An Automated Activity Identification Method for Passively Collected Gps Data', (Ed.)^(Eds.), *3rd Conference on Innovations in Travel Modeling. Phoenix, AZ*.

Ermagun, A., Y. Fan, J. Wolfson, G. Adomavicius and K. Das. (2017), 'Real-Time Trip Purpose Prediction Using Online Location-Based Search and Discovery Services', *Transportation Research Part C: Emerging Technologies* Vol. 77, pp. 96-112.

Ester, M., H.-P. Kriegel, J. Sander and X. Xu. (1996), 'A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise', (Ed.)^(Eds.), *Kdd*.

Feng, T. and H. J. Timmermans. (2016), 'Comparison of Advanced Imputation Algorithms for Detection of Transportation Mode and Activity Episode Using Gps Data', *Transportation Planning and Technology* Vol. 39, No. 2, pp. 180-194.

Flanagin, A. J. and M. J. Metzger. (2008), 'The Credibility of Volunteered Geographic Information', *GeoJournal* Vol. 72, No. 3-4, pp. 137-148.

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

Gilbert, E. and K. Karahalios. (2009), 'Predicting Tie Strength with Social Media', (Ed.)^(Eds.), *Proceedings of the SIGCHI conference on human factors in computing systems*, ACM.

Gong, H., C. Chen, E. Bialostozky and C. T. Lawson. (2012), 'A Gps/Gis Method for Travel Mode Detection in New York City', *Computers, Environment and Urban Systems* Vol. 36, No. 2, pp. 131-139.

Gong, L., R. Kanamori and T. Yamamoto. (2017), 'Data Selection in Machine Learning for Identifying Trip Purposes and Travel Modes from Longitudinal Gps Data Collection Lasting for Seasons', *Travel Behaviour and Society*.

Gong, L., T. Morikawa, T. Yamamoto and H. Sato. (2014), 'Deriving Personal Trip Data from Gps Data: A Literature Review on the Existing Methodologies', *Procedia-Social and Behavioral Sciences* Vol. 138, pp. 557-565.

Gonzalez, M. C., C. A. Hidalgo and A.-L. Barabasi. (2008), 'Understanding Individual Human Mobility Patterns', *nature* Vol. 453, No. 7196, pp. 779.

Gonzalez, P. A., J. S. Weinstein, S. J. Barbeau, M. A. Labrador, P. L. Winters, N. L. Georggi and R. Perez. (2010), 'Automating Mode Detection for Travel Behaviour Analysis by Using Global Positioning Systems-Enabled Mobile Phones and Neural Networks', *IET Intelligent Transport Systems* Vol. 4, No. 1, pp. 37-49.

Griffin, T. and Y. Huang. (2005), 'A Decision Tree Classification Model to Automate Trip Purpose Derivation', (Ed.)^(Eds.), *The Proceedings of the ISCA 18th International Conference on Computer Applications in Industry and Engineering*, Citeseer.

Gu, Y., Y. Yao, W. Liu and J. Song. (2016), 'We Know Where You Are: Home Location Identification in Location-Based Social Networks', (Ed.)^(Eds.), *Computer Communication and Networks (ICCCN), 2016 25th International Conference on*, IEEE.

Guido, G., A. Vitale, V. Astarita, F. Saccomanno, V. P. Giofré and V. Gallelli. (2012), 'Estimation of Safety Performance Measures from Smartphone Sensors', *Procedia-Social and Behavioral Sciences* Vol. 54, pp. 1095-1103.

Géron, A. (2017), *Hands-on Machine Learning with Scikit-Learn and Tensorflow: Concepts, Tools, and Techniques to Build Intelligent Systems*, " O'Reilly Media, Inc.".

Hard, E. N., B. T. Chigoy, S. P. Farnsworth and L. L. Green. (2017), 'Comparison of Cell, Gps, and Bluetooth Derived External Od 1 Data–Results from the 2014 Tyler, Texas Study 2', *technology* Vol. 37, pp. 38.

Horak, R. (2007), *Telecommunications and Data Communications Handbook*, Wiley Online Library.

Huang, L., Q. Li and Y. Yue. (2010), 'Activity Identification from Gps Trajectories Using Spatial Temporal Pois' Attractiveness', (Ed.)^(Eds.), *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on location based social networks*, ACM.

Itsubo, S. and E. Hato. (2006), 'Effectiveness of Household Travel Survey Using Gps-Equipped Cell Phones and Web Diary: Comparative Study with Paper-Based Travel Survey', (Ed.)^(Eds.).

Kattan, L. and B. Abdulhai. (2006), 'Noniterative Approach to Dynamic Traffic Origin-Destination Estimation with Parallel Evolutionary Algorithms', *Transportation Research Record: Journal of the Transportation Research Board*, No. 1964, pp. 201-210.

Kim, Y., F. C. Pereira, F. Zhao, A. Ghorpade, P. C. Zegras and M. Ben-Akiva. (2014), 'Activity Recognition for a Smartphone Based Travel Survey Based on Cross-User History Data', (Ed.)^(Eds.), *Pattern Recognition (ICPR), 2014 22nd International Conference on*, IEEE.

Kohla, B. and M. Meschik. (2013), 'Comparing Trip Diaries with Gps Tracking: Results of a Comprehensive Austrian Study', *Transport survey methods: best practice for decision making*, pp. 305-320.

Kosinski, M., D. Stillwell and T. Graepel. (2013), 'Private Traits and Attributes Are Predictable from Digital Records of Human Behavior', *Proceedings of the National Academy of Sciences* Vol. 110, No. 15, pp. 5802-5805.

Krygsman, S. C. and J. Nel. (2009), 'The Use of Global Positioning Devices in Travel Surveys-a Developing Country Application', *SATC 2009*.

Leber, J. (2013), 'How Wireless Carriers Are Monetizing Your Movements', *MIT Technology Rev*.

Liao, Y., W. Lam, S. Jameel, S. Schockaert and X. Xie. (2016), 'Who Wants to Join Me?: Companion Recommendation in Location Based Social Networks', (Ed.)^(Eds.), *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, ACM.

Liu, F., D. Janssens, G. Wets and M. Cools. (2013), 'Annotating Mobile Phone Location Data with Activity Purposes Using Machine Learning Algorithms', *Expert Systems with Applications* Vol. 40, No. 8, pp. 3299-3311.

Liu, H. X. and W. Ma. (2008), 'Virtual Probe Approach for Time-Dependent Arterial Travel Time Estimation', (Ed.)^(Eds.).

Liu, Q., S. Wu, L. Wang and T. Tan. (2016), 'Predicting the Next Location: A Recurrent Model with Spatial and Temporal Contexts', (Ed.)^(Eds.), *AAAI*.

Lu, L., Y. Xu, C. Antoniou and M. Ben-Akiva. (2015), 'An Enhanced Spsa Algorithm for the Calibration of Dynamic Traffic Assignment Models', *Transportation Research Part C: Emerging Technologies* Vol. 51, pp. 149-166.

Lu, X. and E. I. Pas. (1999), 'Socio-Demographics, Activity Participation and Travel Behavior', *Transportation Research part A: policy and practice* Vol. 33, No. 1, pp. 1-18.

Lu, Y. and L. Zhang. (2015), 'Imputing Trip Purposes for Long-Distance Travel', *Transportation* Vol. 42, No. 4, pp. 581-595.

Ma, J., H. Dong and H. Zhang. (2007), 'Calibration of Microsimulation with Heuristic Optimization Methods', *Transportation Research Record: Journal of the Transportation Research Board*, No. 1999, pp. 208-217.

Marchal, P., S. Roux, S. Yuan, J. Hubert, J. Armoogum, J. Madre and M. Lee-Gosselin. (2008), 'A Study of Non-Response in the Gps Sub-Sample of the French National Travel Survey 2007–08', (Ed.)^(Eds.), *Proceedings of the 8th International Conference on Survey Methods in Transport, France*.

McCulloch, W. S. and W. Pitts. (1990), 'A Logical Calculus of the Ideas Immanent in Nervous Activity', *Bulletin of mathematical biology* Vol. 52, No. 1-2, pp. 99-115.

McGowen, P. and M. McNally. (2007), 'Evaluating the Potential to Predict Activity Types from Gps and Gis Data', (Ed.)^(Eds.), *Transportation Research Board 86th Annual Meeting, Washington*, Citeseer.

Montini, L., N. Rieser-Schüssler, A. Horni and K. Axhausen. (2014), 'Trip Purpose Identification from Gps Tracks', *Transportation Research Record: Journal of the Transportation Research Board*, No. 2405, pp. 16-23.

Moro, A., B. Garbinato and V. Chavez-Demoulin. (2018), 'Discovering Demographic Data of Users from the Evolution of Their Spatio-Temporal Entropy', *arXiv preprint arXiv:1803.04240*.

Murakami, E. and D. P. Wagner. (1999), 'Can Using Global Positioning System (Gps) Improve Trip Reporting?', *Transportation research part c: emerging technologies* Vol. 7, No. 2-3, pp. 149-165.

Naaman, M. (2011), 'Geographic Information from Georeferenced Social Media Data', *SIGSPATIAL Special* Vol. 3, No. 2, pp. 54-61.

Nitsche, P., P. Widhalm, S. Breuss, N. Brändle and P. Maurer. (2014), 'Supporting Large-Scale Travel Surveys with Smartphones–a Practical Approach', *Transportation Research Part C: Emerging Technologies* Vol. 43, pp. 212-221.

Oliveira, M., P. Vovsha, J. Wolf and M. Mitchell. (2014), 'Evaluation of Two Methods for Identifying Trip Purpose in Gps-Based Household Travel Surveys', *Transportation Research Record: Journal of the Transportation Research Board*, No. 2405, pp. 33-41.

Papinski, D., D. M. Scott and S. T. Doherty. (2009), 'Exploring the Route Choice Decision-Making Process: A Comparison of Planned and Observed Routes Obtained Using Person-Based Gps', *Transportation research part F: traffic psychology and behaviour* Vol. 12, No. 4, pp. 347-358.

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

Pearson, D. (2001), 'Global Positioning System (Gps) and Travel Surveys: Results from the 1997 Austin Household Survey', (Ed.)^(Eds.), *Eighth Conference on the Application of Transportation Planning Methods, Corpus Christi, Texas*.

Pourret, O., P. Naïm and B. Marcot. (2008), *Bayesian Networks: A Practical Guide to Applications*, John Wiley & Sons.

Quinlan, J. R. (1986), 'Induction of Decision Trees', *Machine learning* Vol. 1, No. 1, pp. 81-106.

Rasmussen, T. K., J. B. Ingvardson, K. Halldórsdóttir and O. A. Nielsen. (2013), 'Using Wearable Gps Devices in Travel Surveys: A Case Study in the Greater Copenhagen Area', (Ed.)^(Eds.), *Proceedings from the Annual Transport Conference at Aalborg University) ISSN*.

Riederer, C., Y. Kim, A. Chaintreau, N. Korula and S. Lattanzi. (2016), 'Linking Users across Domains with Location Data: Theory and Validation', (Ed.)^(Eds.), *Proceedings of the 25th International Conference on World Wide Web*, International World Wide Web Conferences Steering Committee.

Riederer, C. J., S. Zimmeck, C. Phanord, A. Chaintreau and S. M. Bellovin. (2015), 'I Don't Have a Photograph, but You Can Have My Footprints.: Revealing the Demographics of Location Data', (Ed.)^(Eds.), *Proceedings of the 2015 ACM on Conference on Online Social Networks*, ACM.

Roy, A. and E. Pebesma. (2017), 'A Machine Learning Approach to Demographic Prediction Using Geohashes', (Ed.)^(Eds.), *Proceedings of the 2nd International Workshop on Social Sensing*, ACM.

Schuessler, N. and K. Axhausen. (2009), 'Processing Raw Data from Global Positioning Systems without Additional Information', *Transportation Research Record: Journal of the Transportation Research Board*, No. 2105, pp. 28-36.

Schönfelder, S., K. W. Axhausen, N. Antille and M. Bierlaire. (2002), 'Exploring the Potentials of Automatically Collected Gps Data for Travel Behaviour Analysis', *Arbeitsberichte Verkehrs-und Raumplanung* Vol. 124.

Sermons, M. W. and F. S. Koppelman. (1996), 'Use of Vehicle Positioning Data for Arterial Incident Detection', *Transportation Research Part C: Emerging Technologies* Vol. 4, No. 2, pp. 87-96.

Shen, L. and P. R. Stopher. (2013), 'A Process for Trip Purpose Imputation from Global Positioning System Data', *Transportation Research Part C: Emerging Technologies* Vol. 36, pp. 261-267.

Shen, L. and P. R. Stopher. (2014), 'Review of Gps Travel Survey and Gps Data-Processing Methods', *Transport Reviews* Vol. 34, No. 3, pp. 316-334.

Solomon, A., A. Bar, C. Yanai, B. Shapira and L. Rokach. (2018), 'Predict Demographic Information Using Word2vec on Spatial Trajectories', (Ed.)^(Eds.), *Proceedings of the 26th conference on user Modeling, adaptation and personalization*.

Song, C., T. Koren, P. Wang and A.-L. Barabási. (2010), 'Modelling the Scaling Properties of Human Mobility', *Nature Physics* Vol. 6, No. 10, pp. 818.

Song, C., Z. Qu, N. Blumm and A.-L. Barabási. (2010), 'Limits of Predictability in Human Mobility', *Science* Vol. 327, No. 5968, pp. 1018-1021.

Soto, V., V. Frias-Martinez, J. Virseda and E. Frias-Martinez. (2011), 'Prediction of Socioeconomic Levels Using Cell Phone Records', (Ed.)^(Eds.), *International Conference on User Modeling, Adaptation, and Personalization*, Springer.

Spall, J. C. (1992), 'Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation', *IEEE transactions on automatic control* Vol. 37, No. 3, pp. 332-341.

Spall, J. C. (1998), 'An Overview of the Simultaneous Perturbation Method for Efficient Optimization', *Johns Hopkins apl technical digest* Vol. 19, No. 4, pp. 482-492.

Spall, J. C. (2005), *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*, John Wiley & Sons.

Stopher, P., E. Clifford, J. Zhang and C. FitzGerald. (2008), 'Deducing Mode and Purpose from Gps Data', *Institute of Transport and Logistics Studies*, pp. 1-13.

Stopher, P., C. FitzGerald and J. Zhang. (2008), 'Search for a Global Positioning System Device to Measure Person Travel', *Transportation Research Part C: Emerging Technologies* Vol. 16, No. 3, pp. 350-369.

Stopher, P. and L. Wargelin. (2010), 'Conducting a Household Travel Survey with Gps: Reports on a Pilot Study', (Ed.)^(Eds.), *12th World Conference on Transport Research*.

Stopher, P. R., C. J. Moutou and W. Liu. (2013), 'Sustainability of Voluntary Travel Behaviour Change Initiatives: A 5-Year Study'.

Sui, D. Z. and M. F. Goodchild. (2001), 'Gis as Media?', *International Journal of Geographical Information Science* Vol. 15, No. 5, pp. 387-390.

Troped, P. J., M. S. Oliveira, C. E. Matthews, E. K. Cromley, S. J. Melly and B. A. Craig. (2008), 'Prediction of Activity Mode with Global Positioning System and Accelerometer Data', *Medicine and science in sports and exercise* Vol. 40, No. 5, pp. 972-978.

Tsui, S. and A. Shalaby. (2006), 'Enhanced System for Link and Mode Identification for Personal Travel Surveys Based on Global Positioning Systems', *Transportation Research Record: Journal of the Transportation Research Board*, No. 1972, pp. 38-45.

Tympakianaki, A., H. N. Koutsopoulos and E. Jenelius. (2015), 'C-Spsa: Cluster-Wise Simultaneous Perturbation Stochastic Approximation Algorithm and Its Application to Dynamic Origin–Destination Matrix Estimation', *Transportation Research Part C: Emerging Technologies* Vol. 55, pp. 231-245.

Wagner, D. (1997), 'Lexington Area Travel Data Collection Test: Gps for Personal Travel Surveys', *Final Report, Office of Highway Policy Information and Office of Technology Applications, Federal Highway Administration, Battelle Transport Division, Columbus*, pp. 1-92.

Wagner, D., J. Seymour, N. Malcosky, J. Kinateder and J. Nguyen. (1998), 'Determining Heavy Duty Truck Activity Using Global Positioning System (Gps) Technology', (Ed.)^(Eds.), *8th CRC On-Road Vehicle Emissions Workshop, San Diego*.

Wang, P., J. Guo, Y. Lan, J. Xu and X. Cheng. (2016), 'Multi-Task Representation Learning for Demographic Prediction', (Ed.)^(Eds.), *European Conference on Information Retrieval*, Springer.

Wen, C.-H. and F. S. Koppelman. (2001), 'The Generalized Nested Logit Model', *Transportation Research Part B: Methodological* Vol. 35, No. 7, pp. 627-641.

Wiehe, S. E., A. E. Carroll, G. C. Liu, K. L. Haberkorn, S. C. Hoch, J. S. Wilson and J. Fortenberry. (2008), 'Using Gps-Enabled Cell Phones to Track the Travel Patterns of Adolescents', *International journal of health geographics* Vol. 7, No. 1, pp. 22.

Wolf, J., R. Guensler and W. Bachman. (2001), 'Elimination of the Travel Diary: Experiment to Derive Trip Purpose from Global Positioning System Travel Data', *Transportation Research Record: Journal of the Transportation Research Board*, No. 1768, pp. 125-134.

Wolf, J., R. Guensler, L. Frank and J. Ogle. (2000), 'The Use of Electronic Travel Diaries and Vehicle Instrumentation Packages in the Year 2000 Atlanta Regional Household Travel Survey: Test Results, Package Configurations, and Deployment Plans', (Ed.)^(Eds.), *9th International Association of Travel Behaviour Research Conference*.

Wolf, J., S. Schönfelder, U. Samaga, M. Oliveira and K. Axhausen. (2004), 'Eighty Weeks of Global Positioning System Traces: Approaches to Enriching Trip Information', *Transportation Research Record: Journal of the Transportation Research Board*, No. 1870, pp. 46-54.

Wolf, J. L. (2000), 'Using Gps Data Loggers to Replace Travel Diaries in the Collection of Travel Data', Citeseer.

Work, D. B., O.-P. Tossavainen, Q. Jacobson and A. M. Bayen. (2009), 'Lagrangian Sensing: Traffic Estimation with Mobile Devices', (Ed.)^(Eds.), *American Control Conference, 2009. ACC'09.*, IEEE.

Xiao, G., Z. Juan and C. Zhang. (2016), 'Detecting Trip Purposes from Smartphone-Based Travel Surveys with Artificial Neural Networks and Particle Swarm Optimization', *Transportation Research Part C: Emerging Technologies* Vol. 71, pp. 447-463.

Xiao, Y., D. Low, T. Bandara, P. Pathak, H. B. Lim, D. Goyal, J. Santos, C. Cottrill, F. Pereira and C. Zegras. (2012), 'Transportation Activity Analysis Using Smartphones', (Ed.)^(Eds.), *Consumer Communications and Networking Conference (CCNC), 2012 IEEE*, IEEE.

Yalamanchili, L., R. Pendyala, N. Prabaharan and P. Chakravarthy. (1999), 'Analysis of Global Positioning System-Based Data Collection Methods for Capturing Multistop Trip-Chaining Behavior', *Transportation Research Record: Journal of the Transportation Research Board*, No. 1660, pp. 58-65.

Ying, J. J.-C., Y.-J. Chang, C.-M. Huang and V. S. Tseng. (2012), 'Demographic Prediction Based on Users Mobile Behaviors', *Mobile Data Challenge*, pp. 1-6.

Zhang, L., S. Dalyot, D. Eggert and M. Sester. (2011), 'Multi-Stage Approach to Travel-Mode Segmentation and Classification of Gps Traces', *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences:[Geospatial Data Infrastructure: From Data Acquisition And Updating To Smarter Services] 38-4 (2011), Nr. W25* Vol. 38, No. W25, pp. 87-93.

Zhang, L. and Y. Lu. (2012), 'Innovative Data Collection and Modeling Methods for Long-Distance Passenger Travel Demand Analysis', (Ed.)^(Eds.).

Zhang, L. and Y. Lu. (2015), 'Us National and Inter-Regional Travel Demand Analysis: Person-Level Microsimulation Model and Application to High-Speed Rail Demand Forecasting', (Ed.)^(Eds.).

Zhao, Y. (2000), 'Mobile Phone Location Determination and Its Impact on Intelligent Transportation Systems', *IEEE Transactions on intelligent transportation systems* Vol. 1, No. 1, pp. 55-64.

Zhong, E., B. Tan, K. Mo and Q. Yang. (2013), 'User Demographics Prediction Based on Mobile Data', *Pervasive and mobile computing* Vol. 9, No. 6, pp. 823-837.

Zhong, Y., N. J. Yuan, W. Zhong, F. Zhang and X. Xie. (2015), 'You Are Where You Go: Inferring Demographic Attributes from Location Check-Ins', (Ed.)^(Eds.), *Proceedings of the eighth ACM international conference on web search and data mining*, ACM.

Zhou, C., P. Ludford, D. Frankowski and L. Terveen. (2005), 'An Experiment in Discovering Personally Meaningful Places from Location Data', (Ed.)^(Eds.), *CHI'05 Extended Abstracts on Human Factors in Computing Systems*, ACM.

Zito, R., G. D'este and M. Taylor. (1995), 'Global Positioning Systems in the Time Domain: How Useful a Tool for Intelligent Vehicle-Highway Systems?', *Transportation Research Part C: Emerging Technologies* Vol. 3, No. 4, pp. 193-209.

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