U.S. Department of Transportation

Federal Highway Administration

Next Generation National Household Travel Survey National Origin Destination Data

Passenger Origin-Destination Data Methodology Documentation

UNIVERSITY OF MARYLAND

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Next Generation National Household Travel Survey National Origin Destination Data Passenger Origin-Destination Data Methodology Documentation		5. Report Date October 2021 Revise date: April 2023 6. Performing Organization Code:
7. Author/s Lei Zhang, Aref Darzi, Yixuan Pan, Mofeng Yang, Qianqian Sun, Aliakbar Kabiri, Guangchen Zhao, Mohammad Ashoori, Saeed Saleh Namadi, Chenfeng Xiong, Gregory W. Jordan, Kathleen Stewart, Michael L. Pack		8. Performing Organization Report No.
9. Performing Organization Name and Address University of Maryland Office of Research Administration 3112 Lee Building College Park, MD 20742-5141		10. Work Unit No. 11. Contract or Grant No. GS-10F-0502N
12. Sponsoring Organization Name and Address Office of Highway Policy Information Federal Highway Administration 200 New Jersey Avenue, SE Washington, DC 20590		13. Type of Report and Period Covered Draft Methods Document. October 2020–April 2021 14. Sponsoring Agency Code
15. Supplementary Notes The Federal Highway Administration's Task Monitors for this project were Daniel Jenkins, P.E. and Dr. Patrick Zhang, P.E.		
16. Abstract This document outlines methods and approaches to develop origin destination (OD) passenger data covering all aspects of data collection, analysis, and final tabulation. This document includes sample representative analysis, tour and trip identification, modal identification methods, trip purpose identification mechanisms, population expansion approaches, etc.		
17. Key Words Origin destination passenger data, passively collected location data, travel behavior, emerging technologies, data privacy, National Household Travel Survey (NHTS), NextGen NHTS	18. Distribution Statement: Internal draft. NOT for public distribution.	

19. Security Classification (of this report) Unclassified	20. Security Classification (of this page) Unclassified	21. No. Of Pages	22. Price N/A

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized.

EXECUTIVE SUMMARY

Federal Highway Administration (FHWA) launched the Next Generation National Household Travel Survey (NextGen NHTS) program with the goal of establishing a more continuous national travel monitoring program. The program includes the production of national multimodal passenger and truck travel Origin-Destination (OD) tables from passively collected mobile device location data.

This document describes the technical approach employed by the University of Maryland (UMD) project team to develop national passenger OD data for the program. The methodology for national truck OD production is documented in a separate deliverable. For the production of the national passenger OD product, the team employs a tour-based approach to properly identify all tours and trips from passively collected data, including trip origin, destination, start time, and end time. For each identified trip, imputation algorithms are then applied to produce travel mode and trip purpose, and trip distance is derived. Devices and trips are expanded based on control totals at various levels. A national expanded all-trip roster is obtained for the development of OD data products at the national level. Key methodology highlights include:

1) The UMD team receives data from multiple providers of passively collected passenger travel data.
2) The team compiles the source data and establishes a national raw data panel with more than 20 standardized quality metrics.
3) A tour-based approach is employed to properly recognize tours, linked trips, unlinked trips, and intermediate stops.
4) A series of validated imputation algorithms are used to identify home/work locations, trip purposes, trip distances, and travel modes.
5) A multi-level data expansion process is applied to address various types of sampling biases at both device and trip levels.

In addition, the team has developed a rigorous validation plan for the proposed algorithms and the final data products at both individual and aggregate levels to ensure high product quality of the national passenger OD data. The UMD team establishes product validation targets based on the NHTS core survey, National Transit Database (NTD), Airline Origin and Destination Survey (DB1B), Air Carrier Statistics Database (T-100), Highway Performance Monitoring System (HPMS), and other available datasets.

The team is fully committed to enhancing the transparency of both the raw data and methodological steps in producing the national passenger OD products. This document reports the technical approach and validation plan. The team publishes all quality metrics of the 2020 raw data set in this document. National passenger and truck OD products are published by FHWA in the public domain, with access to the associated source codes for the computation algorithms used in the development of these products available upon request.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. OVERVIEW OF THE TECHNICAL APPROACH AND METHODOLOGY

This document describes the technical approach employed by the University of Maryland (UMD) team to develop high-quality national passenger Origin-Destination (OD) data for the Next Generation National Household Travel Survey (NextGen NHTS) OD Data Program.

Figure 1 provides an overview of the overall methodology. The "National Device and Location Data Panel Construction" first preprocessed raw sighting data from multiple passenger data sources. Raw sighting data quality was evaluated based on sample size, representativeness, sighting data accuracy, data frequency, data consistency, and other quality metrics. Additional steps were performed to assemble the national data panel, including home and fixed workplace identification, device deduplication, and sighting data integration. After the national device and location data panel was constructed, a tour-based approach was employed to properly process the data and identify all tours and trips from the raw location data, including trip origin, destination, start time, and end time. For each identified trip, imputation algorithms were then applied to produce travel mode and trip purpose, and trip distance was derived. The result is a "national all-trip roster", which was stored in a trip roster format for the development of the national passenger OD data product.

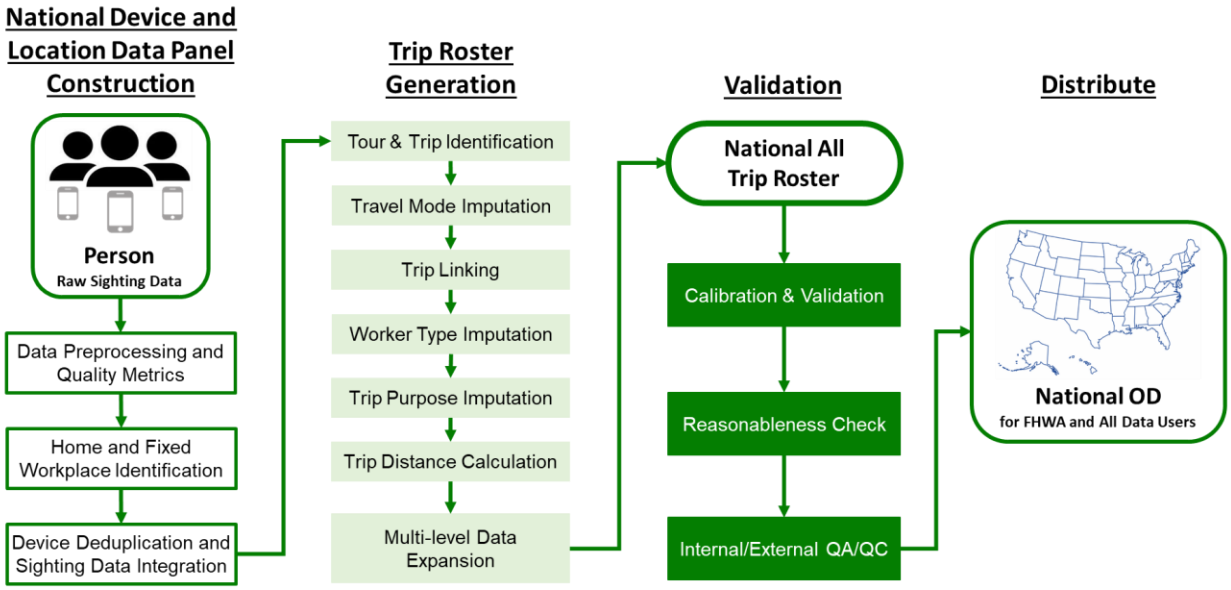


Figure 1. Passenger OD data production flow chart for the Next Generation National Household Travel Survey (NextGen NHTS) OD Data Program

The entire national all-trip roster was used by the UMD team to develop national passenger OD products. Trips were expanded based on population and employment data, imputed socio-demographics, and a multi-level data expansion method that employed expansion factors at both mobile device and trip levels. In the "Validate" step, the UMD team calibrated and validated OD products based on the 2017 National Household Travel Survey (NHTS), the 2020 Traffic Volume Trends (TVT) reports, the 2020 National Transit Database (NTD), the 2020 Airline Origin and

Destination Survey (DB1B) data, the 2020 Air Carrier Statistics Database (T-100), , and other validation data. Before the “Distribute” step, which delivered national OD data products for FHWA and all data users, a rigorous quality assurance and quality control (QAQC) procedure was implemented by an internal UMD check and an external and independent assessment.

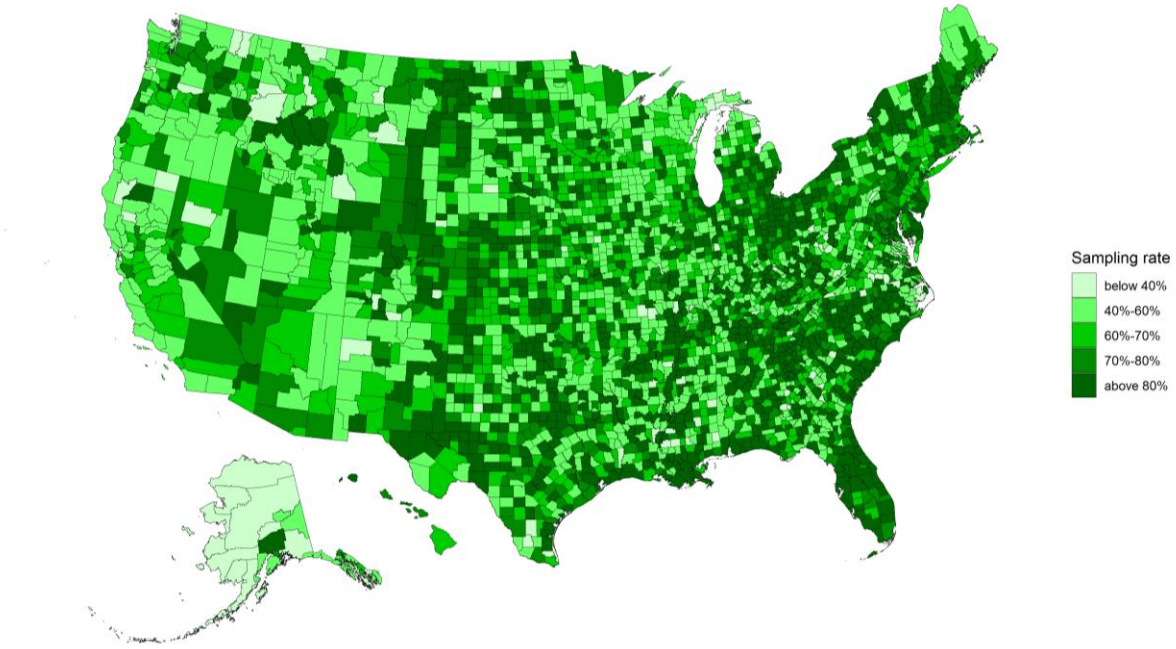
2. RAW SIGHTING DATA ASSEMBLY, PREPROCESSING, AND QUALITY EVALUATION

This section describes the methodology for assessing raw location data quality, identifying the home and workplace information for each device, deduplicating devices, and creating the device and location data panel for future trip-level information imputation.

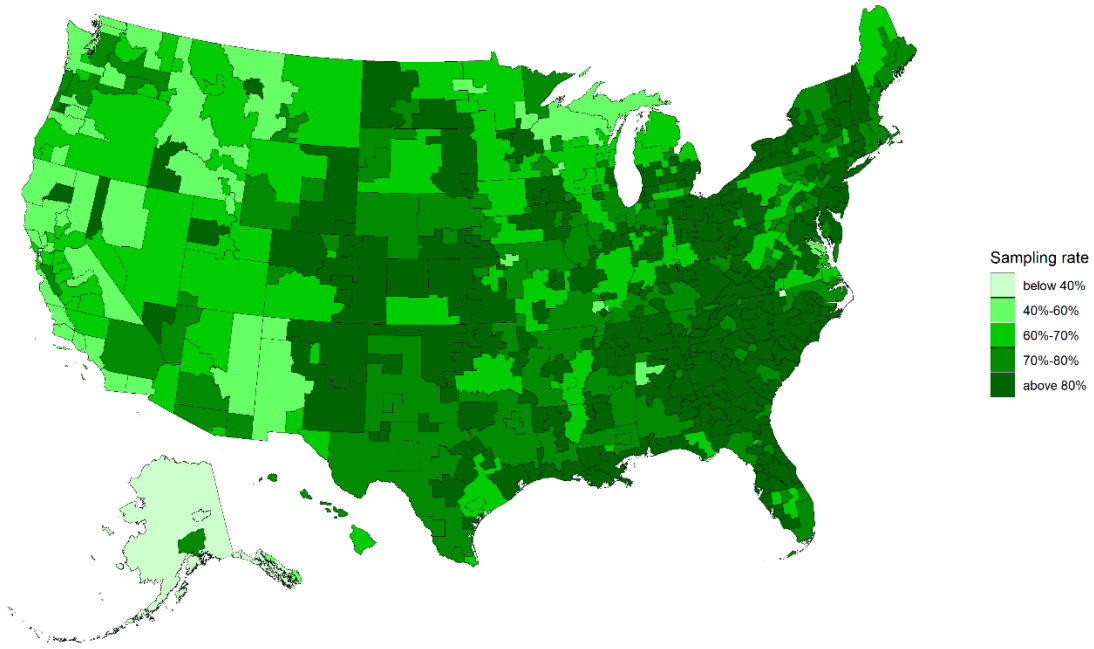
2.1. Data Preprocessing and Quality Metrics

Passively collected mobile device location data generated from various positioning technologies such as cellphone, Global Positioning System (GPS), and location-based services (LBS), have become increasingly available for transportation planning and operations. A location sighting is generated when a mobile application updates the device's location with the most accurate sources based on existing location sensors such as Wi-Fi, Bluetooth, cellular tower, or GPS (Chen et al., 2016; Wang and Chen, 2018). The location sighting can reflect the exact location of mobile devices and thus provide location information describing individual-level mobility patterns. Typically, one location sighting includes an anonymized device identifier (ID), latitude and longitude coordinates, time stamps, positioning accuracy, etc. Such location data will be referred to as sighting or sighting data in the remaining document.

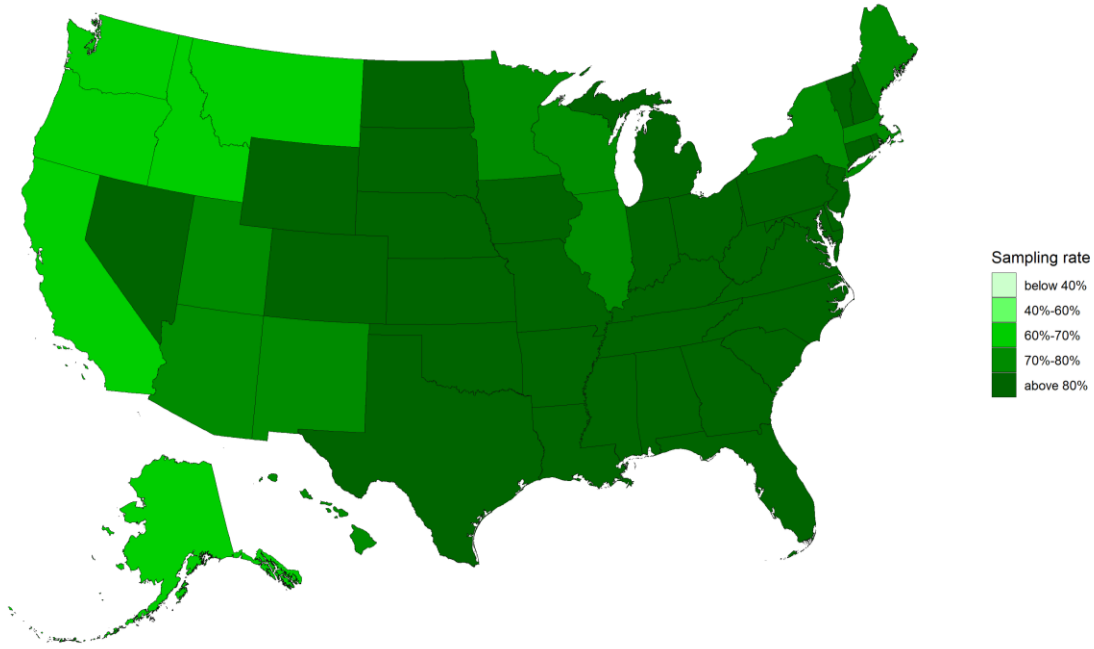
The UMD team developed a cloud-storage based method to ingest raw location sightings from multiple data vendors and form the raw sighting data panel. For the 2020 NextGen NHTS passenger OD data product, the raw sighting data panel consisted of more than 270,000,000 Monthly Active Users (MAU) and represented movements across the nation. Figure 2 depicts the coverage of the raw sighting data at different geographical levels.



(a) Sampling rate at county level



(b) Sampling rate at MSA level



(c) Sampling rate at state level

Figure 2. Raw sampling rate of raw sighting data employed in this project (a) at the county level, (b) at the MSA level, and (c) at the state level for 2020 passenger OD data product

Various dimensions of assessing data quality, such as consistency, accuracy, completeness, and timeliness, were discussed in the literature (e.g., Cappiello et al., 2003; Batini et al., 2006; Wang

and Chen, 2018) and in the team's previous work (Zhang et al., 2020). A comprehensive framework that assessed the raw sighting data quality from the four dimensions, addressed the quality issues through data preprocessing, and evaluated the cleaned sighting data using quality metrics is shown in Figure 3. The details on data preprocessing and quality metrics are given in Sections 2.1.1 and 2.1.2.

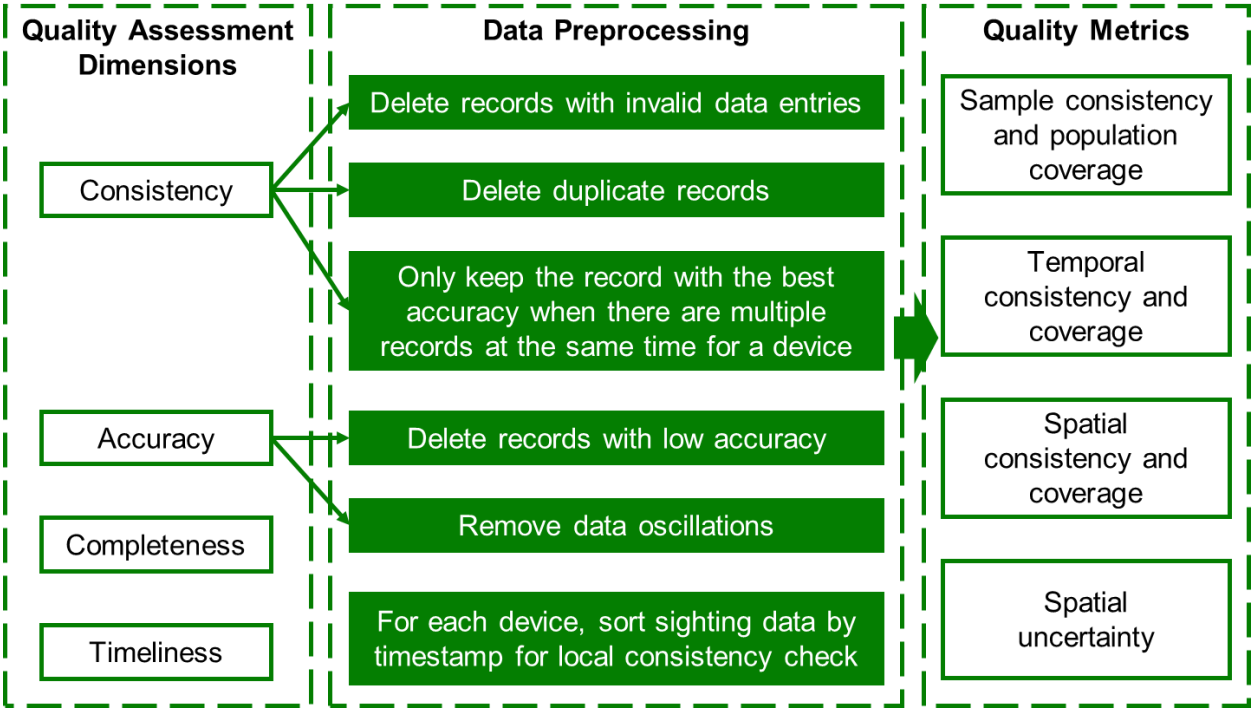


Figure 3. Schema of the data quality evaluation and data preprocessing

2.1.1. Data Preprocessing

Raw sighting data were preprocessed separately for each data provider. The data preprocessing includes the following steps.

- **Step 1**: remove raw sightings with invalid data entries, e.g., negative values for latitudes.
- **Step 2**: remove duplicate sightings considering all data attributes.
- **Step 3**: clean multiple sightings with the same timestamp for the same device. Based on the ranking of location accuracy, the sighting with the smallest location uncertainty is reserved.
- **Step 4**: remove raw sightings with low location accuracy (defined as greater than 492 feet (150 meters)), a threshold selected based on a sensitivity analysis evaluating the trade-off between location uncertainty and percentage of sightings removed.
- **Step 5:** identify and remove data oscillations.
- **Step 6:** for each device, sort the sightings by timestamps.

The procedure of removing data oscillations (Step 5) is summarized in Figure 4. Data oscillations are abnormal movements with unreasonable distance and time combinations between sightings. They exist in the raw sighting data due to known and unknown technical errors that occur during the data collection process. To simplify the extraction of moving patterns of devices and increase the computation efficiency, device trajectories were denoted by a sequence of level-7 geohash zones instead of latitudes and longitudes. Geohash is a public domain geocode system that encodes a geographic location into a short string of letters and digits. There are twelve levels of geohash zones, which differ in zone size, length of the zone name, etc. Specifically, the level-6 geohash zones (i.e., a grid of about 4000 × 2000 feet) and level-7 geohash zones (i.e., a grid of about 500 × 500 feet) were utilized in the current and following data processing steps. The simplified trajectories were utilized for detecting oscillations.

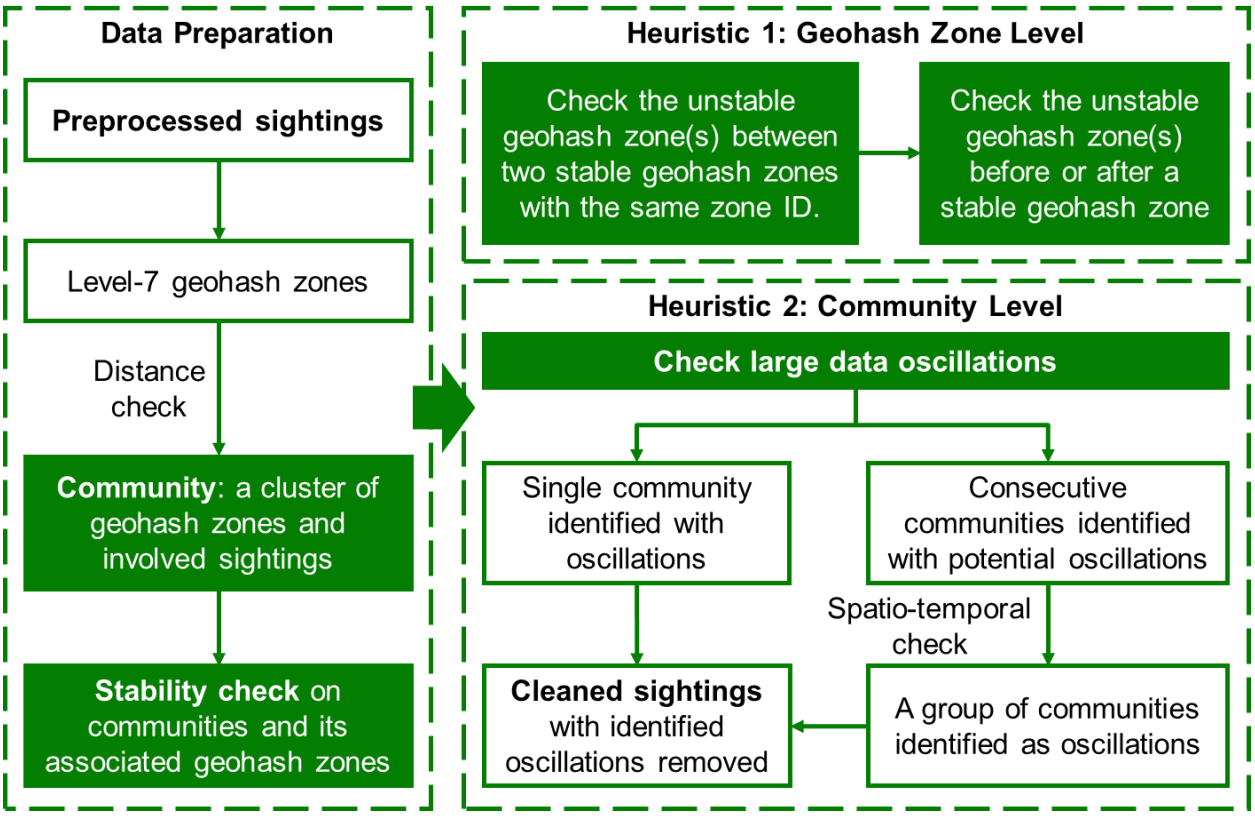


Figure 4. Procedure for removing data oscillations

If a device was observed within a community (i.e., within a specific location range smaller than 0.5 mile) frequently enough (with more than 5 sightings) or long enough (for more than 5 minutes), the corresponding sightings were treated as true visits and form a "stable community." Based on the identified true visits, other locations were investigated to check oscillations. All level-7 geohash zones involved in a stable community were determined to be stable level-7 geohash zones. Communities and level-7 geohash zones were used to remove oscillations in different cases.

Two heuristic rules were designed to remove oscillations:

- Heuristic 1 at the geohash zone level: if a device left a stable level-7 geohash zone and returned to the same zone within 30 seconds, the sightings out of the stable zone during the 30 seconds were determined to be oscillations and were removed. In addition, if a device moved more than 5 miles away from a stable level-7 geohash zone to an unstable level-7 geohash zone in 2.5 minutes, all sightings in that unstable level-7 geohash zone were determined to be oscillations and are removed.
- Heuristic 2 at the community level (Figure 5): (a) between two nearby communities, C1 and C3, if the device moved to a faraway community C2 at high speed, the corresponding sightings in C2 were removed; (b) if the device moved at high speed between two groups of communities—the odd communities (C1, C3, and C5) and the even communities (C2 and C4)—the group of communities with shorter dwell time were considered oscillations and their corresponding sightings were removed. Specifically, the nearby and faraway communities were relative positions decided by the spatial-temporal criteria from Heuristic 1. The criteria utilized the two intercommunity speeds between C1-C2 and C2-C3, the three intercommunity distances between C1-C2, C2-C3, and C1-C3, and the dwell time of the middle community C2. When scenario (a) continuously happened, such as the continuously unstable communities, C1, C2, …, and C5, shown in Figure 5 (b), the dwell time of each group of communities was used for reserving one group of communities as the stable communities.

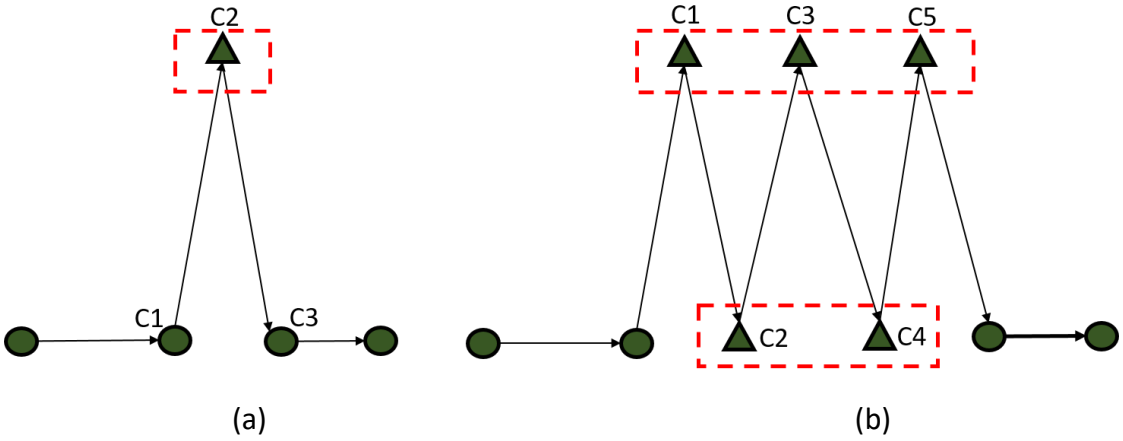


Figure 5. Two scenarios of data oscillations considered by Heuristic 2

2.1.2. Data Quality Metrics

A set of quality metrics was employed to assess the preprocessed sighting data from each data provider. The high quality of sighting data contributes to a better representation of the entire population and a better coverage of each device's movements. The essential metrics employed in this project included sample consistency and population coverage (i.e., monthly active users, daily active users, and regularly active users), temporal consistency and coverage (i.e., temporal consistency, data frequency, device representativeness, active local hours, hourly coverage, and daily coverage), spatial consistency and coverage (i.e., geographical representativeness), and spatial uncertainty (i.e., location accuracy). The definition of each metric is described along with

the corresponding statistics for each metric was summarized in Table 1. All metrics in Table 1 were derived from one-month of raw sighting data panel in 2020. The values are provided to help data users compare the data quality of this raw sighting data panel with that of other similar data sources.

Table 1. Definition and descriptive statistics of the data quality metrics

Quality Metric	Definition	2020 One-Month Raw Sighting Panel	Interpretation
Monthly Active Users (MAU) ***(in devices)***	The number of devices with at least one sighting for a specific month	270,601,232	Implies a sampling rate of more than 80% on a monthly basis
Daily Active Users (DAU) ***(in devices on average)***	The number of devices with at least one sighting on a specific day for a specific month	112,420,233	Implies an average sampling rate of about 34% on a daily basis
Regularly Active Users (RAU) ***(in devices)***	The number of devices with at least seven days of more than ten daily sightings for a specific month	68,016,290	Indicates a sampling rate greater than 20% regarding temporally consistent devices
Temporal Consistency ***(in days)***	The average number of observed days for RAUs in a specific month.	24.2 (max limit possible = 31 days)	Indicates the level of temporal consistency and coverage of the RAUs
Data Frequency ***(in sightings)***	Mean, 25^{th}, 50^{th} , and 75^{th} percentile of the average daily number of sightings by RAU devices	Mean = 234.4 25^{th} = 72.4 50^{th} = 127.8 75^{th} = 298.2	Indicates the sighting frequency of RAUs
Location Accuracy ***(in feet)***	Mean, 25^{th}, 50^{th} , and 75^{th} percentile of the positioning accuracy of RAU devices. Positioning accuracy is defined as the maximum distance between a device's recorded location and its actual location at 95% confidence level	Mean = 49.2 25^{th} = 13.1 50^{th} = 31.1 75^{th} = 64.6	Indicates the reliability of location sightings of RAUs
Geographical Representativeness ***(by devices)***	Variance of population coverage among different counties, measured by a Gini coefficient[1] between 0 and 1, with 0 indicating equal sampling rate in all zones and 1 indicating that all RAUs are from a single zone	Gini = 0.4	Indicates an even geographical distribution of RAUs per population
Geographical Representativeness ***(by sighting)***	Variance of sighting volume divided by county-level population, measured by a Gini coefficient between 0 and 1, with 0 indicating equal sighting volume per person in all zones and 1 indicating that all sightings are from a single zone	Gini = 0.2	Indicates an even geographical distribution of sightings per population

[1] Gini coefficient (Gini, 1912) is a statistical measure of the equality of a given data. It can be calculated by the ratio of the area above the Lorenz curve to the summation of the area above and the area below the Lorenz curve. The Lorenz curve is a graph showing the distribution of the given data.

Device Representativeness ***(by average daily sighting volume)***	Variance in the average daily number of sightings among RAU devices, measured by a Gini coefficient between 0 and 1, with 0 indicating equal sighting frequency and 1 indicating distinct sighting frequency for all RAUs	Gini = 0.6	Indicates a notable uneven distribution of average daily sighting volume for each RAU, which may be a result of distinct smartphone use behaviors and travel behaviors of different device owners. A data expansion framework was developed to address the uneven distribution.
Active Local Hours ***(in hours)***	Mean, 25^{th}, 50^{th} , and 75^{th} percentile of the average daily number of local hours observed for RAUs	Mean = 6.4 25^{th} = 2.3 50^{th} = 4.8 75^{th} = 8.9	Indicates a high temporal consistency and coverage of the RAUs
Hourly Coverage ***(by average hourly sighting)***	Variance in the average sighting volume by the hour of the day for all RAUs, measured by a Gini coefficient between 0 and 1, with 0 indicating an equal average number of sightings from the 24 hours and 1 indicating all sightings are from one hour	Gini = 0.2	Indicates an even distribution of average daily number of sightings among the 24 hours for RAUs
Daily Coverage ***(by total daily sighting)***	Variance in the total sighting volume by day of the month for all RAUs, measured by a Gini coefficient between 0 and 1, with 0 indicating an equal total number of sightings from each day in one month and 1 indicating all sightings are from one day	Gini = 0.1	Indicates an even distribution of daily total number of sightings across all days in the month for RAUs

2.2. The Identification of Home and Fixed Workplace

Due to privacy protection, the upstream data providers or data vendors anonymize all the sample devices from the sighting data. This means the sighting data generally does not contain any personal information, such as home location, age group, or income level. Such personal information is critical in sample bias correction and data expansion. For the national passenger OD data development, the framework only used sighting data from sample devices whose home locations could be imputed. The sample devices with imputed home locations were further distinguished as devices with fixed workplaces (the fixed workplace is different from home), devices without fixed workplaces but with jobs, and devices without fixed workplaces or jobs based on the mobility patterns (Pan et al., 2023).

The UMD team first employed a behavior-based method to identify the home and fixed workplace location based on the cleaned sighting data (see Section 2.1) and further imputed more socio-demographic information using machine learning methods after identifying trip-level information. The behavior-based method evaluated the temporal patterns of places observed for every device and ranks the frequently visited locations to identify the home and fixed workplace.

Samples with identified home but without fixed workplace might have occupations like transportation and shipping occupations, whose trips were covered in the national truck OD data products, and cleaning and home maintenance workers, whose trips were still considered in the national passenger OD data product and whose working profiles were necessary for device-level expansion. Those occupations without fixed workplaces generally induce more driving trips than others. Therefore, an additional step considered the spatio-temporal patterns of their driving trips and imputed their worker type to facilitate the device-level expansion and ensure the proper coverage of the national passenger OD data product (see Section 3.4). Those unemployed and those who work from home were categorized as devices without fixed workplace or jobs by the algorithm since there was a lack of evidence to distinguish between the two types.

Home and fixed workplace identification are built upon activity location identification, i.e., identifying the most significant locations for each device from a set of activity locations. For Call Detail Record (CDR) data, one location record corresponds to one cell tower, and the covered area of an observed cell tower is intuitively defined as an activity location. For sighting data generated from cellular data and location-based services (LBS), the sightings include latitudes and longitudes. Therefore, a clustering method is typically applied, with the centroid of the cluster identified as an activity location. After identifying the activity locations, the next step is to impute the type of activity conducted in each place as either home or fixed workplace.

There are two types of methods for the imputation of activity type: behavior-based and context-based (Chen et al., 2016). The behavior-based method infers the home and workplaces based on the most frequently visited places during night and daytime (Phithakkitnukoon et al., 2010; Alexander et al., 2015), or on sighting volume and sighting regularity (Chen et al., 2014). The context-based method considers the surroundings, such as land use and nearby points of interest (POIs), and infers the activity types with empirical rules (Xie et al., 2009; Huang et al., 2010). As

the most widely used and the most applicable method, the behavior-based approach is efficient in determining daily life centers, such as home and workplace, especially when there is a lack of additional personal information in the raw data. The team followed the general idea of the behavior-based approach in developing the framework for imputing home and fixed workplace locations.

Figure 6 introduces the methodology to impute home and fixed workplace locations and worker types.

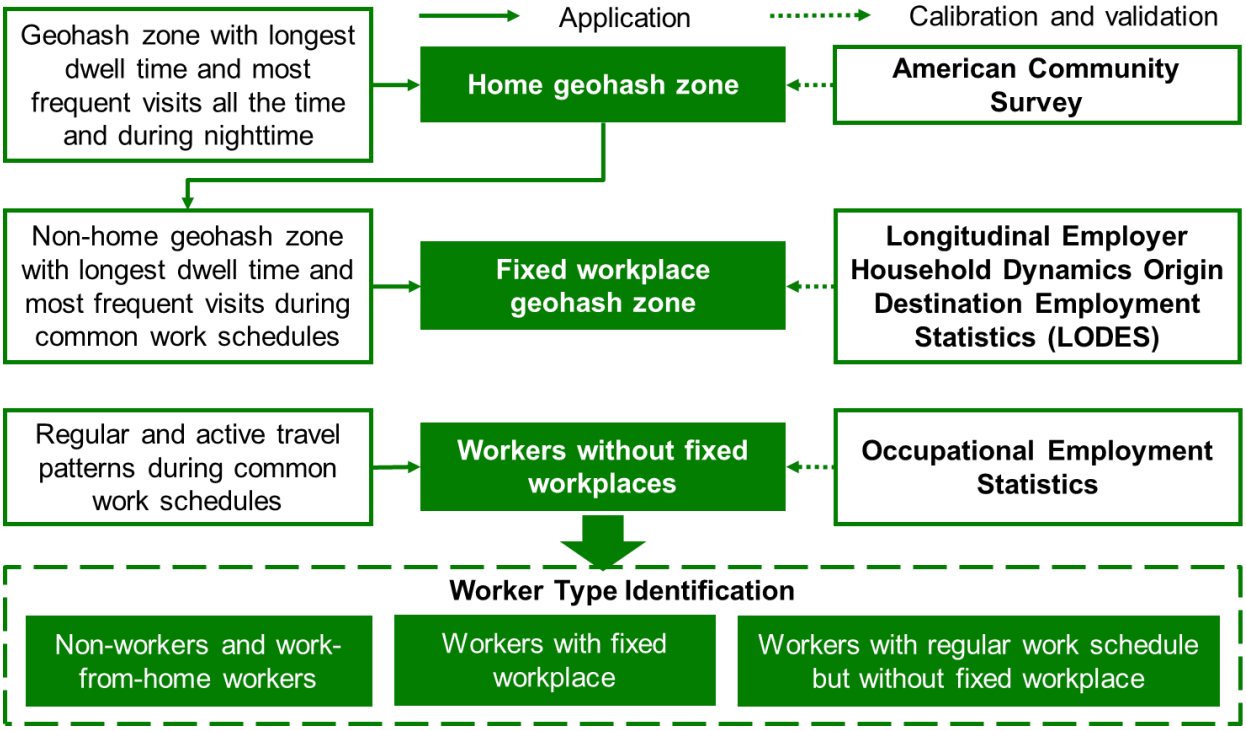


Figure 6. The framework for home, fixed work locations, and worker type imputation

2.2.1. Home Location Identification

To efficiently process the tremendous amount of mobile device location data, the algorithm utilized geohash to aggregate the latitudes and longitudes into candidates for activity locations. Considering the location uncertainty of sightings and the possible household activities conducted around the home, the algorithm first identified the home and workplace at a level-6 geohash zone and then selected the most frequently observed location at a level-7 geohash zone within the identified level-6 geohash zone as a more precise representation of home and fixed workplace.

People spend most of their time, especially nighttime, at home and some fixed and regular hours during daytime at the workplace. The framework first identified three frequently observed level-

6 geohash zones as home location candidates based on the overall observed days in a month (at least three days or half of the total observed days for each device), the average observed hours in those observed days (at least two hours), and the average sightings in those observed hours. The method favored the home location candidate that was most frequently observed during nighttime and selected it as the home location at level-6 geohash zone level. The first two steps were then repeated at a smaller geospatial resolution (level-7 geohash zone) to find a more precise representation of home location. To properly identify nighttime period, the team investigated 2017, 2018, and 2019 American Time Use Survey (ATUS) and defined nighttime as 9:00 p.m.–5:59 a.m., since more than 80% of full-time and part-time workers were observed to visit home at least once during that period.

The parameter for the minimum average number of observed hours, i.e., 2 hours, was calibrated based on the Pearson correlation test between the county-level number of imputed residents and a population over 16 reported by the American Community Survey (ACS) for home location identification. The Pearson correlation value based on the selected parameter was higher than 0.95.

2.2.2. Fixed Workplace Location Identification

With home location identified, the framework recognized an individual's major work location that is not home. Similar to the home location identification, the method considered workplace candidates based on the visiting frequency (at least three workdays, or half of the total observed workdays for each device) and average duration (at least two hours) during daytime on workdays. On top of that, the algorithm introduced a temporal similarity ratio between the workplace candidates and identified home location. The motivation was two-fold. First, for the sake of computation efficiency, the home and workplace imputation adopted geohash as the representation of the actual location. If a device dwelled around the borders of geohash zones, it could be frequently and alternately observed in one or more neighboring geohash zones—twin zones—despite high location accuracy. Such twin zones could outperform the actual workplace zone with regard to visiting frequency, duration, and regularity and thus be misidentified as the workplace. Second, although a minimum commute distance threshold would be an intuitive alternative to partially address the issue, it might compromise workplaces that are close to one's home location. Based on the assumption that individuals commute from home to workplace and work for consecutive hours before commuting back home, the home and workplace are not typically both observed in the same hour. Hence, workplace identification checked the temporal similarity in terms of the specific hours when the device was flagged to be at the identified home location and workplace candidates to find the most possible workplace location.

For each workplace candidate, the temporal similarity ratio was defined as the ratio between the number of hours when the device was flagged to be at both home and the workplace candidate and the number of total hours when the device was flagged to be at the workplace candidate. In an ideal situation where the daily location observations are complete for one device with a fixed workplace, the ratio should be $\frac{2}{Number\ of\ daily\ work\ hours}$ (approximately 0.25) when the commute time is shorter than one hour, and zero when the commute time is longer than one

hour, since the device would not be flagged to be at home and workplace in the same hour. For example, if the device left home at 8:10 AM, arrived at work at 8:50 AM, worked from 9:00 AM to 6:00 PM, left work at 6:05 PM, and arrived home at 6:50 PM, the device was flagged to be at both home and workplace in the hour of 8:00-8:59 AM and the hour of 6:00-6:59 PM, and its number of daily work hours is 11 hours (including the two hours when the device was also flagged to be at home). The similarity ratio would be 0.18. However, most devices would not have complete location observations throughout the month, which is the time window of home and workplace imputation. To address this, the algorithm was designed to favor work candidates with small temporal similarity ratios while imposing a maximum temporal similarity threshold (selected as 0.6) to exclude the inefficient large ratios in distinguishing between the actual workplace zone and the twin zones of home location (Pan et al., 2023).

The parameters for the minimum average number of observed hours, i.e., two hours, were calibrated based on the Pearson correlation test between the county-level number of imputed commuters and the number of workers reported by Longitudinal Employer Household Dynamics (LEHD) Origin Destination Employment Statistics (LODES) for workplace imputation. The maximum temporal similarity threshold was set to be 0.6 for two reasons. First, the workplace should be observed for at least one specific hour when the home was not observed in addition to the potential two shared observed hours during the two commute trips. Second, a Pearson correlation analysis was conducted between the county-level number of imputed commuters and the reported number of workers in LODES. The Pearson correlation value based on the selected parameters was higher than 0.95.

2.3. Device Deduplication and Sighting Data Integration

After identifying the home and fixed workplace for the devices from each data provider, the UMD team developed the algorithm that identified the duplicated devices within and between data providers, integrated the sighting data for the duplicated devices, integrated the device and sighting data from all data providers, and created the national device and sighting data panel for passenger trip identification. Figure 7 illustrates the general steps for creating a high-quality and consistent device and sighting data panel. More details are described in the remainder of the section.

To ensure data quality, devices had to meet at least two out of three predefined criteria in terms of device-level data quality metrics. The three criteria were:

- The average number of sightings per device per day throughout the entire month (at least six observations)
- The number of days that a device was observed in a month (at least 10 days)
- The average number of unique hours daily that a device was observed (at least eight hours)

To ensure the minimum population coverage and avoid privacy concerns for each zone, the effective sampling rate in all the U.S. counties was preferred to be over 5%. Otherwise, all the

devices were kept in the counties with sampling rate lower than 5% to avoid certain biases from very few devices.

This approach was used to construct the initial data panel for the first month of the 2020 OD data product. In order to maintain a consistent device and sighting data panel for the following months, the methodology was modified to keep the maximum number of existing devices in the panel and maintain or improve the panel quality. In the second month, the devices were divided into two groups: devices existing in the previous month's panel and the remaining devices. The devices existing in the previous panel were favored and thus evaluated with relaxed thresholds. If their device-level quality metrics were higher than the relaxed thresholds, they were kept in the data panel. The remaining devices were evaluated against the initial thresholds. This approach was repeated with each new month of data to ensure a high-quality and consistent data panel throughout the entire year.

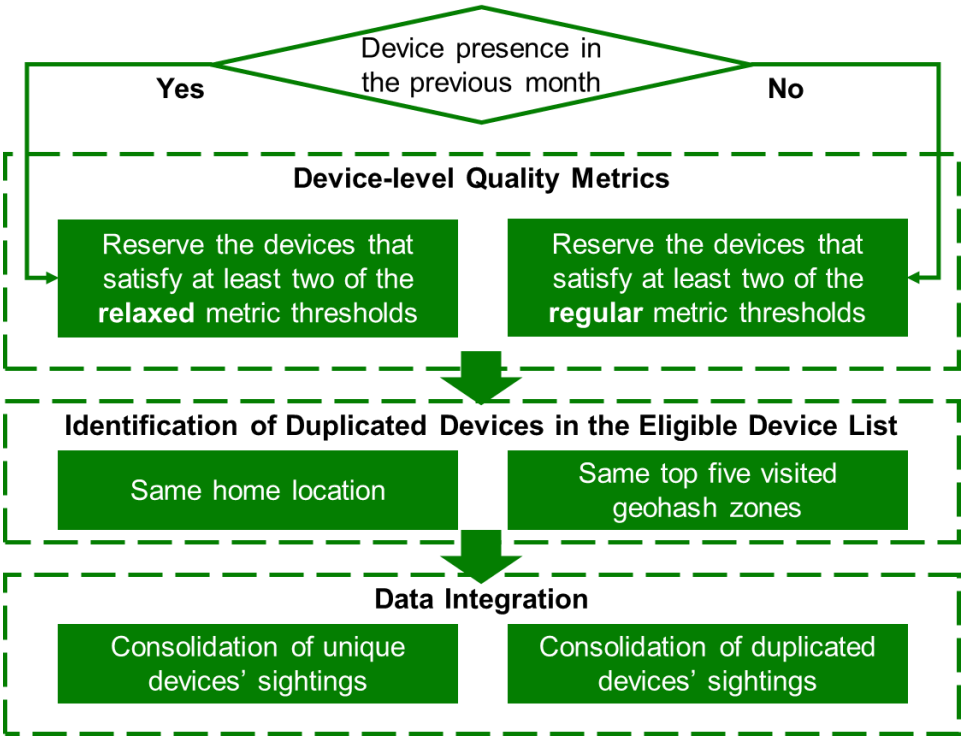


Figure 7. Flowchart of device deduplication and sighting data integration

As personal electronics become more accessible, one could own multiple mobile devices (e.g., smartphones, tablets, and smartwatches), recording one's sighting data and sharing such data with mobile device data vendors. Therefore, an individual's movement may be captured by more than one device in the sighting data. In addition, the sightings from the same device may also be counted more than once when combining multiple sighting datasets to create a more representative and comprehensive device and sighting data panel. To avoid the overrepresentation of individuals owning multiple devices and sharing data with multiple data vendors, a deduplication method was developed to identify the devices that represent the same individuals, i.e., duplicated devices.

To identify duplicated devices integrated from different data providers, two heuristic rules were defined:

- The duplicated devices must have the same imputed home location
- The duplicated devices should share the same top five frequently visited locations within one month

The home location identification algorithm was described in detail in Section 2.2.1. Regarding the second rule, the locations visited by each device were ranked by the total number of unique hours observed and the total number of location observations during a month. Devices that shared the same home location and the same top five most frequently visited locations (which may include home locations as well) were considered duplicated devices. This was a conservative algorithm, which ensured that the actual duplicated devices would be captured but carried a slight risk that could result in some distinct devices being identified as duplicates.

Finally, the sightings of all identified duplicated device IDs were consolidated to provide more reliable and complete trajectories for those devices in the data panel.

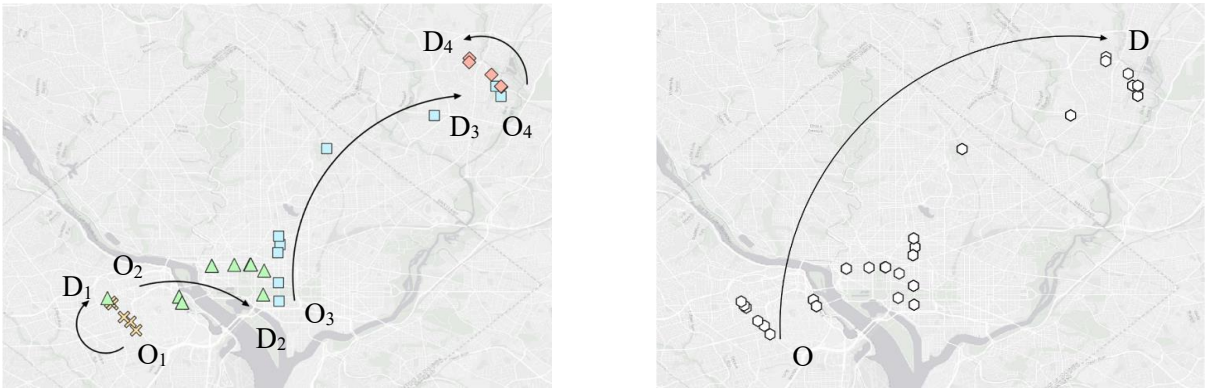
In summary, this Chapter presented the methodology for data preprocessing, quality control, and home and fixed workplace imputation. With these methodological steps, the raw sighting data panel was cleaned and filtered to form the national device and location data panel. The national device and location data panel only included the sample devices with home locations imputed by the proposed methodological framework. As shown in Figure 1, the national device and location data panel was a key input to the national trip roster generation, which are described in detail in Chapters 3 and 4.

3. NATIONAL PASSENGER TRIP DATA DEVELOPMENT

This section describes the methodology for identifying trips, imputing travel mode, linking selected trips, excluding non-passenger trips, imputing trip purpose, and deriving trip distance to create the national all-trip roster after obtaining the national device and location data panel.

3.1. Tour and Trip Identification

Trips are the unit of analysis for almost all transportation applications. Sightings from mobile device location data do not directly include trip information. Therefore, trip identification algorithms were used to extract trip information from the cleaned sightings. The team used a tour-based method to first identify tours and improve the completeness of identified trips. Figure 8 illustrates how the tour-based method produced more accurate trip identification results. Figure 8 (a) and (b) show how the tour-based method differentiated true activity clusters (e.g., home cluster and work cluster) from mid-trip transfer points (e.g., waiting at a transit station). It should be noted that the tour-based approach was also necessary to identify the true origins and destinations of long-distance trips, especially air trips.



(a). Multiple unlinked person trips (b). One linked person home-to-work trip

Figure 8. Tour identification and trip linking demonstration

The following subsections describe the steps for identifying tours and trips. The algorithm was applied to the observations from each device independent of those from other devices.

3.1.1. Home-Based Tour Identification

The algorithm started with each device's identified home location (see Section 2.2.1). The home-based tour identification processed a device's locations every day, from 4 a.m.—4 a.m. the next day, or the "trip day." All the sightings between two at-home observations were considered as a home-based tour. Long-distance tours were defined as tours in which a device was observed equal to or more than 50 miles away from its home location. To be consistent with the majority of reported travel in the core travel survey, it was assumed that unless the device was on a long-distance tour, the device started and ended the trip day at home. In the next step, the sightings of each device were separated into two groups: sightings on short-distance tours and sightings on long-distance tours. Finally, short-distance tours underwent a daily short-distance trip

identification process and long-distance tours went through a monthly long-distance trip identification process.

3.1.2. Trip Identification for Short-Distance Tours

Given that the data included stationary points, a recursive algorithm based on the decision tree model was utilized to identify if the sighting was stationary or moving. The decision tree considered six attributes, i.e., the great circle distance, time interval, and speed between the current sighting and the previous and next sightings. The decision tree had three hyper-parameters: a distance threshold of 984 ft (i.e., 300 meters), a time threshold of 5 minutes, and a speed threshold of 3 miles per hour (3 mph or 1.4 m/s). The speed threshold was used to identify if a sighting was recorded on the move, and the distance and time thresholds was used to identify trip ends.

The recursive algorithm checked every sighting to identify if they started a new trip or belonged to the same trip as the previous sighting (Figure 9). If the previous sighting was not on a trip (i.e., a stationary sighting), the current sighting started a trip if it had a speed faster than 3 mph to the next sighting. If the previous sighting was on a trip, the following rules were checked to identify if the current sighting belonged to the same trip, stopped the trip, or started a new trip:

- If a sighting had a speed greater than 3 mph from the previous sighting, the sighting belonged to the same trip as its previous sighting.
- If a sighting had a speed slower than 3 mph from the previous sighting and was more than 984 ft away from the previous sighting, the sighting did not belong to the same trip as its previous sighting. If the speed to the next sighting was also slower than 3 mph, the current sighting simply terminated the trip; otherwise, it became the start of a new trip.
- If a sighting had a speed slower than 3 mph from the previous sighting was within 984 ft from the previous sighting, the cumulative dwell time for all the consecutive sightings meeting such criteria was computed and checked: 1) if the cumulative dwell time was less than five minutes, the current sighting belonged to the same trip, 2) otherwise, it terminated the trip if the speed to the next sighting was slower than 3 mph or started a new trip if the speed to the next sighting was faster than 3 mph.

The algorithm could identify a local movement as a trip if the device moved within a stay location. To filter out such trips, all trips shorter than 984 ft were removed.

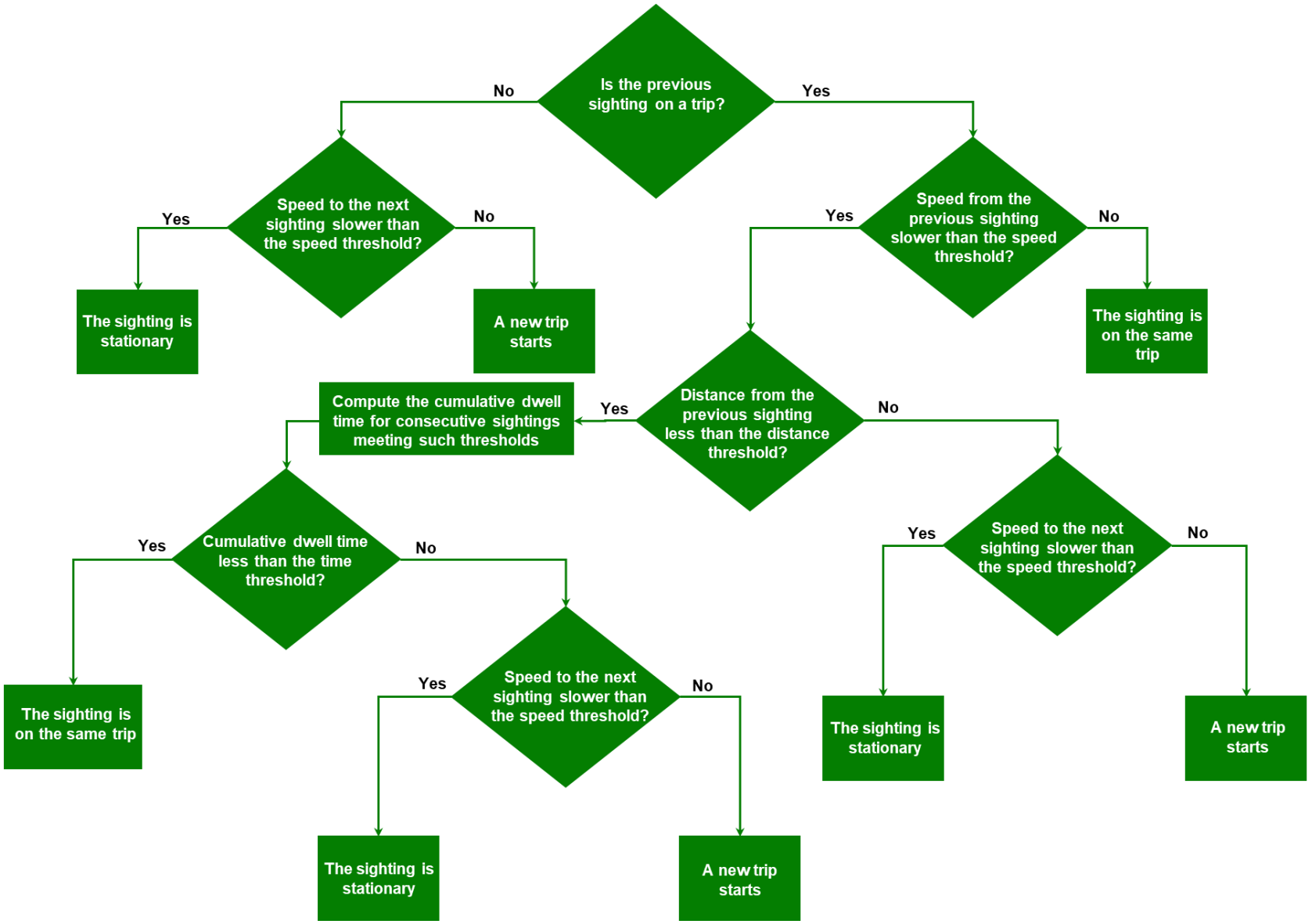


Figure 9. Recursive algorithm for trip identification for short-distance tours

3.1.3. Trip Identification for Long-Distance Tours

Trip identification for long-distance tours followed a different procedure due to the different nature of long-distance trips. To start, all device sightings on long-distance tours for the entire month were filtered. Figure 10 shows the process for identifying long distance tours. Each stage of the flowchart is described in the following subsections.

3.1.3.1. Stop and primary destination identification

A recursive trip identification algorithm, similar to that described in Section 3.1.2, was applied, but with a larger time threshold of 30 minutes instead of 5 minutes, meaning that a trip ended only if the device stayed somewhere for more than 30 minutes. In this step, all the trip ends were identified and named as "secondary stops." Primary stops were then identified from the secondary stops. Primary stops on a long-distance tour were places where the device stayed for a significant amount of time and/or from which the device made local trips. In order to identify the primary stops, each secondary stop was checked against the following criteria:

- The duration of stay in the secondary stop was longer than two hours and during the stay, the device exited and reentered the secondary stop
- The duration of stay at a location was longer than 24 hours
- The secondary stop was the home location

Furthermore, the primary destination of a tour was defined as the farthest stop that was located at least 50 miles away from the home location of the device. The primary destination was unique in one long-distance tour and was first identified from the primary stops. If no primary stop fulfilled the requirement, the primary destination was then identified from the secondary stops.

3.1.3.2. Subtour identification

A subtour was considered a segment of a long-distance tour that fell between two primary stops. Therefore, all sightings between two primary stops were considered to be on the same subtour.

3.1.3.3. Trip identification

If a long-distance tour did not have a primary destination or had the same primary destination as the identified workplace, the short-distance trip identification algorithm (with a time threshold of five minutes) was applied to all the sightings in the tour. If a tour had a primary destination different from the fixed workplace, the long-distance trip identification algorithm with a time threshold of 30 minutes was applied to sightings between two different primary stops, and the short-distance trip identification recursive algorithm with a time threshold of 5 minutes was applied to sightings around the same primary stop (local trips around a primary stop on a long-distance tour).

Finally, all the tours, subtours, and trips were stored for the following steps, such as mode imputation, trip linking, trip purpose imputation.

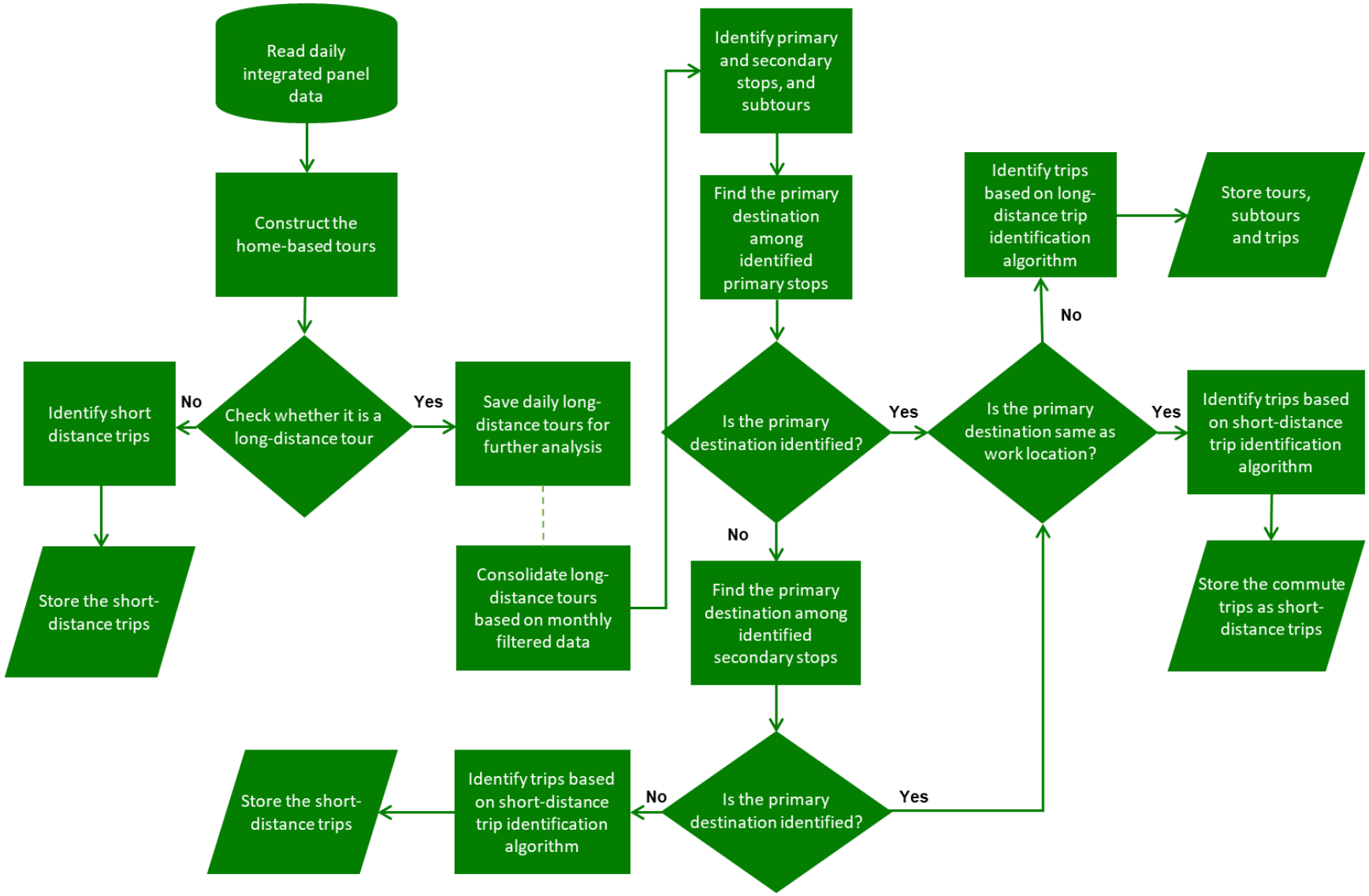


Figure 10. Recursive algorithm for trip identification for long-distance tours

3.2. Travel Mode Imputation

The typical methods and features to impute travel mode from sighting data are summarized as follows:

- Trip-based approach: the trip-based approach is based on already identified trips, where each trip has only one travel mode to be imputed (e.g., Gong et al., 2012).
- Segment-based approach: the segment-based approach separates the sighting data into fixed-length segments in terms of time or distance, and then imputes the travel mode for each segment (e.g., Stenneth et al., 2011). Then the segments with the same travel mode are further merged to form a single-mode trip.

However, when imputing travel mode from the sighting data, one key issue is that the location recording intervals (LRIs) of data from different sources varies significantly. In some cases, the LRI might be high and less information might be captured, which makes it hard to accurately impute the travel mode. To address this issue and as part of the FHWA EAR Pilot Project, *Data Analytics and Modeling Methods for Tracking and Predicting Origin-Destination Travel Trends Based on Mobile Device Data* (Zhang et al., 2020), the UMD team collected sighting data with labeled travel mode information (Yang et al. 2021) via a series of dedicated smartphone studies, accumulating thousands of multimodal samples with ground truth information.

For the NextGen NHTS national passenger OD data product, mode was imputed in stages. The air travel mode was firstly imputed based on a heuristic rule calibrated based on ground truth data. Then, an ensemble machine learning model was developed and used to impute ground transportation travel modes with both the information from the mobile device location data itself and the multimodal transportation network information. Figure 11 shows the flowchart of the travel mode imputation method. More details are presented in the following sections.

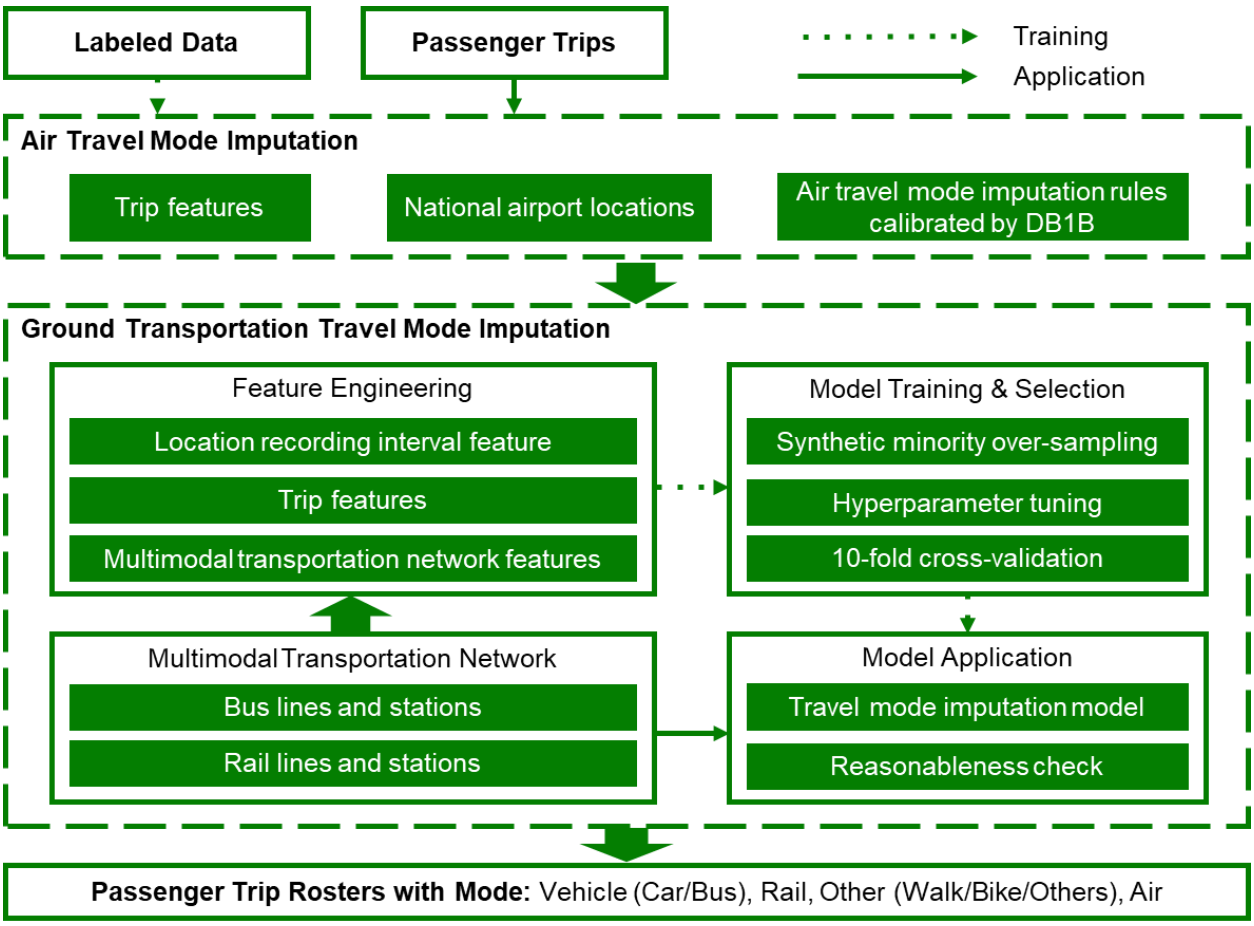


Figure 11. Flowchart of travel mode imputation

3.2.1. Air Travel Mode Imputation

As shown in Figure 11, because of its uniqueness in trip features compared to the ground transportation travel modes, the first step was to impute air trips from the national passenger trip roster. The air trips were extracted by calibrating a heuristic rule with four parameters: (1) travel time, (2) travel distance, (3) the average travel speed, and (4) the origin/destination distances to the nearest airport. The DB1B data was used as the ground truth data to calibrate the aforementioned four parameters in order to maximize the correlation between the number of trips between each airport OD pair identified from mobile device location data and reported from DB1B. The calibrated values of these four parameters are shown below:

- The origin-destination straight-line distance of an air trip was longer than 50 miles
- The travel time of an air trip was longer than 30 minutes
- The average travel speed of an air trip was faster than 75 mph
- The origin and destination distances to the airport were both shorter than two miles (the two-mile threshold is calibrated to achieve the highest OD flow correlation with DB1B)

After identifying the air trips using the four parameters, one additional layer of reasonable ness check was conducted: if the travel time, travel distance, and average travel speed were significantly high and did not belong to any ground transportation mode, the nearest airports were assigned for both origin and destination of the trip.

3.2.2. Ground Transportation Travel Mode Imputation

After air trips were imputed from the national passenger trip roster, a machine learning model was developed and applied to impute the ground transportation travel modes for non-air trips, including vehicle (car and bus), rail, and active transportation and ferry (walk, bike, ferry, and other modes). More details are presented in the following sub-sections.

3.2.2.1. Feature engineering

Feature engineering directly affects the model performances, i.e., imputation accuracy. Three types of features (including a total of 32 variables) were considered for ground transportation travel mode imputation, as shown in Table 2.

Table 2. Features for Detecting Ground Transportation Travel Mode

Features	Number of Variables
Location Recording Interval Feature	
Average # of records per minute	1
Trip Features	
Origin-destination straight-line distance	1
Cumulative trip distance	1
Travel time	1
Average travel speed	1
0^{th}, 5^{th}, 25^{th}, 50^{th}, 75^{th}, 95^{th}, 100^{th} percentile travel speed	7
Multimodal Transportation Network Features	
0^{th}, 5^{th}, 25^{th}, 50^{th}, 75^{th}, 95^{th}, 100^{th} percentile distance to the nearest rail lines	7
0^{th}, 5^{th}, 25^{th}, 50^{th}, 75^{th}, 95^{th}, 100^{th} percentile distance to the nearest bus lines	7
Origin/Destination distances to the nearest rail station	2
Origin/Destination distances to the nearest bus stop	2
Percentage of records within 165-feet of all rail stations	1
Percentage of records within 165-feet of all bus stops	1

The LRI feature, represented by the average number of sightings per minute, indicates the location service usage during a trip. The trip features can show the characteristics of each trip, including the origin-destination straight-line distance, cumulative trip distance (network distance), travel time, average travel speed, and different percentiles of travel speed, which were all derived from the FHWA EAR Pilot Project sighting data. The multimodal transportation network features are important to distinguish between different ground transportation travel modes. Here, the distance for each sighting to its nearest rail and bus lines were generated to calculate the 0^{th}, 5^{th}, 25^{th}, 50^{th}, 75^{th}, 95^{th}, and 100^{th} percentile distance to rail and bus lines; the distance for the origin/destination of each trip to its nearest rail and bus stations/stops were also

calculated. Also, the percentage of records within 165 feet (50 meters) of all rail stations or bus stops were calculated for each trip. Those features were used to capture the short stops at rail or bus stations for rail and bus travels since more sightings would be observed very closely around those stations when people wait for the transit services. However, due to the variations in LRI and location accuracy, the sightings could be observed at a further distance from the stations, which relaxed the distance threshold to 165 feet. The U.S. national bus and rail lines and bus stops and rail stations (including metro and Amtrak Stations) were collected from the Homeland Infrastructure Foundation-Level Data (HIFLD) and U.S. Department of Transportation Bureau of Transportation Statistics.

3.2.2.2. Random forest model and its accuracy

After comparing the performance of different machine learning models, the Random Forest (RF) machine learning model was selected as the final model to impute the ground transportation travel modes. The model was trained using over 11,000 sample data with labeled travel mode information (Yang et al. 2021). Synthetic Minority Over-Sampling Technique (SMOTE) was then applied to the training data to address the imbalanced sample problem, where the minority class from the existing samples was synthesized (Bohte and Maat, 2009). The randomized search approach was used to fine-tune the model. During the model training process, 10-fold cross-validation (CV) was conducted to evaluate the model performance. The training results showed that the RF model could achieve 97.1% cross-validation accuracy for ground transportation travel mode imputation. The trips with the imputed four modes were further aggregated into three modes, including vehicle (car and bus), rail, and active transportation and ferry (walk, bike, ferry, and other modes).

3.3. Merging Unlinked Trip into Linked Trips

All the trips derived from the trip identification step were considered as the unlinked trips. The UMD team developed a trip linking algorithm to examine the characteristics of the unlinked trips (i.e., tour type and the imputed mode). For those unlinked trips that were not merged by this proposed algorithm, they were directly considered as linked trips in the final trip roster.

For short- and long-distance tours, two separate methodologies to link the related unlinked trip segments were developed, as shown in Figure 12.

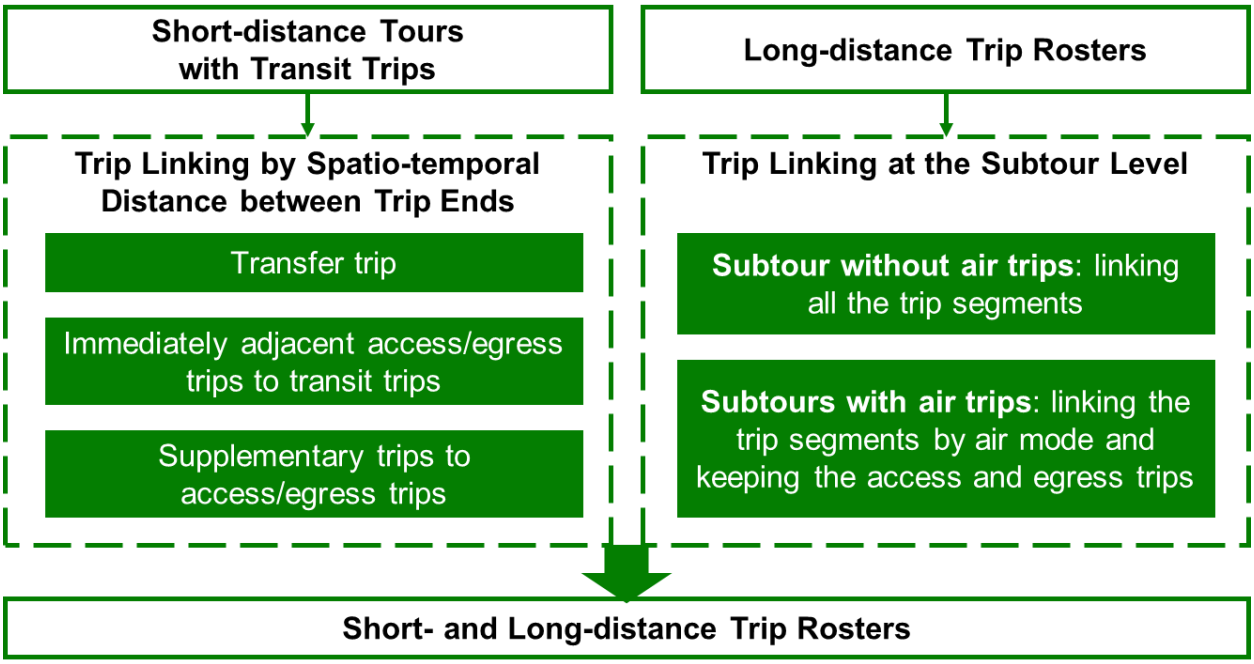


Figure 12. Flowchart of merging unlinked trip segments into trips

For short-distance tours, trip linking was only conducted for transit trips in order to recover the actual travel demand. One linked transit trip could consist of the following six types of unlinked trip segments:

- Access trip to the transit mode: either car mode or active transportation and ferry (ATF) modes (e.g., bike or walk)
- Transit trips: either bus or rail mode
- Same-transit-mode transfers: same mode as its previous transit trip
- Change of transit mode: different mode as its previous transit trip
- Egress trip from the transit mode: either car mode or ATF modes
- Supplementary trip(s) before the access trip
- Supplementary trip(s) after the egress trip

After locating all unlinked transit trips, different spatial and temporal thresholds were used to link different types of segments as follows:

- Linking the access or egress trips to the transit trips: both the distance and the time difference between the trip ends satisfied the spatial and temporal threshold values (at most 0.5 mile and 20 minutes).
- Linking the supplementary trips to the access or the egress trips: one more step was taken to link the trips right before the access trip or right after the egress trip. For example, in the case of park and ride, the actual access trips to the transit mode could consist of both a walking segment(s) and a driving segment(s). Two parameters—the spatial distance (at most 0.2 miles) and time difference (at most 5 minutes) between two trip ends—were checked to make the access and egress trips complete. The spatial and temporal

threshold values applied here were more restrictive, considering the waiting time and possible activity space at the transit stations.

- Linking either same-mode or different-mode transfer trips: the time difference between two transit trip ends were smaller than a transfer time threshold value (at most 30 minutes).

The five spatial or temporal threshold values were calibrated by a series of sensitivity analyses based on two critical ratios: 1) the ratio between the number of unlinked trips related to linked transit trips and the number of linked transit trips, and 2) the ratio between the number of unlinked transit trips and the number of linked transit trips (transit transfer ratio). The selected threshold values resulted in similar values for the two ratios compared with the 2017 NHTS estimates and the transit transfer ratio reported by American Public Transportation Association (APTA) (Clark, 2017).

As for long distance tours, trips between two primary stops were linked, i.e., each subtour was one linked trip unless it includes an air trip. When there is an air trip, this air trip between airports forms one linked trip by itself. The access trips going to the airport were linked as one trip with the major ground transportation mode as the new travel mode and the egress trips leaving from the airport were linked as one trip.

3.4. Worker Type Identification

As described in Section 2.2, the sample devices with imputed home locations were labeled as workers if they also had imputed fixed workplaces. For the remaining devices, the potential workers without fixed workplace were evaluated based on their travel behavior statistics. Therefore, one additional step for worker type identification was conducted following the trip-level information extraction described in Sections 3.1, 3.2, and 3.3.

The worker type identification had two major objectives: 1) to identify and remove trips made by professional drivers so that the passenger trip estimates for the population were exclusive of the trips made by professional drivers driving for work which were captured in the national truck OD data; and 2) to identify other workers without fixed workplaces (the list of occupations is summarized in Section 3.4.2), whose trips were considered in the national passenger OD data products and whose work profiles were necessary for device-level expansion.

3.4.1. Professional Driver Identification

In order to exclude the trips from professional drivers in the national passenger OD data products, an algorithm to first identify professional drivers was applied. The team conducted a practice scan on heuristic algorithms for identifying professional drivers and the trip-level features of those drivers before designing the identification algorithm. A flowchart of this algorithm is shown in Figure 13.

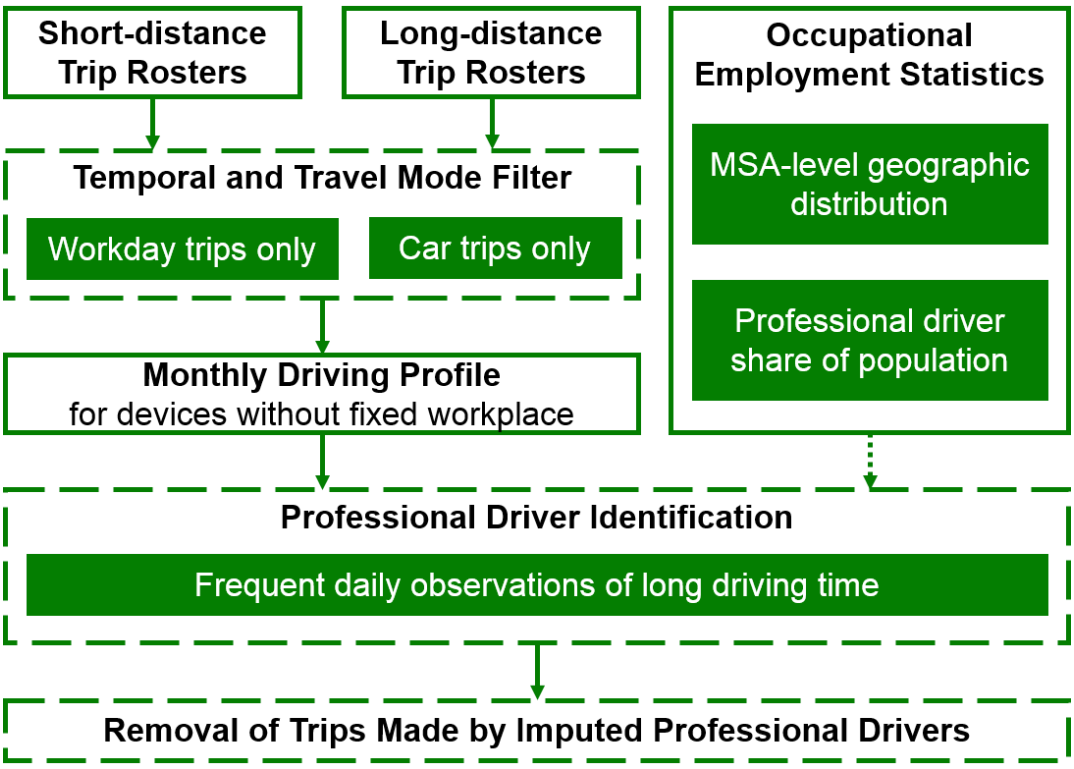


Figure 13. Flowchart of removing professional driver trips

Key features of professional drivers are their driving trips with long trip durations and the regularity of such behavior. According to the U.S. Census Bureau's 2019 American Community Survey, 90-minute or longer one-way commutes account for 3.1% of all commute trips (Burd et al., 2021). The professional drivers' daily travel time were typically higher than 90 minutes. According to the hours of service (HOS) (USDOT FMCSA, 2020), commercial drivers of passengers can drive up to 14 hours followed by at least 10 consecutive hours off duty; commercial drivers of property can drive up to 11 hours followed by at least 10 consecutive hours off duty. Another important feature of professional drivers is that they regularly drive for a long time. Some individuals might also drive for long hours for personal recreation. However, that behavior occasionally happens and usually happens on weekends while the long-hour driving behavior of professional drivers is frequent and can happen every day. According to a sample survey (Hanowski et al., 2001), most truck drivers worked on a Monday–Friday or a Tuesday–Friday schedule. The aforementioned features constituted the basics of the professional driver identification algorithm.

To identify and exclude professional driver trips, the algorithm utilized the percentage of observed workdays with long-time driving behavior (i.e., total driving time in a day is greater than a threshold value). The algorithm used a relaxed criterion that at least 50% of the observed workdays of each device show long-hour driving behavior (i.e., total driving time in a day is more than three hours). Meanwhile, a minimum number of workdays (nine days) was added as another threshold. The parameters for the minimum driving hours (three hours) and the minimum number of workdays (nine workdays) were selected based on the Pearson correlation test

between the MSA-level number of imputed professional drivers and the reported number of professional drivers by the Occupational Employment and Wage Statistics (OEWS).

3.4.2. Other Workers without Fixed Workplaces

After identifying the professional drivers and excluding their trips from the passenger OD data production, the following occupation categories defined by the 2018 Standard Occupational Classification (SOC) system were considered as workers without fixed workplace:

- 33-2021 Fire inspectors and investigators
- 33-2022 Forest fire inspectors and prevention specialists
- 33-3051 Police and sheriff's patrol officers
- 33-3052 Transit and railroad police
- 43-5041 Meter readers, utilities
- 49-9050 Line installers and repairers
- 49-9080 Wind turbine service technicians
- 41-9091 Door-to-door sales workers, news and street vendors, and related workers

The typical travel behavior of such workers without fixed workplaces was frequent and regular travel during the daytime. In the ATUS, more than 25% of the full-time workers were at a workplace between 6:00 a.m.–5:59 p.m. According to the 2018 ACS survey, commuters with commute times longer than 45 minutes were up to 12%. Therefore, one hour of commute time was added to the daytime window (6:00 a.m.–5:59 p.m.). As a result, the time window of 5:00 a.m.–5:59 p.m. was then adopted as the daytime period for identifying workers without fixed workplaces. The workers without fixed workplaces were defined as devices that make more than 5 driving trips longer than 10 minutes away from home on at least 8 workdays or half of the workdays during the month that the device is observed making trips. The parameters for the minimum driving trips (five driving trips) and the minimum number of workdays (eight workdays) were selected based on the Pearson correlation test between the MSA-level number of imputed workers without fixed workplaces and the reported number by the OEWS.

3.5. Trip Purpose Imputation

With worker profiles identified, trip purpose was imputed as work and non-work. The imputation process included two major parts: data preparation and imputation algorithms. Figure 14 shows the flowchart of the trip purpose imputation method.

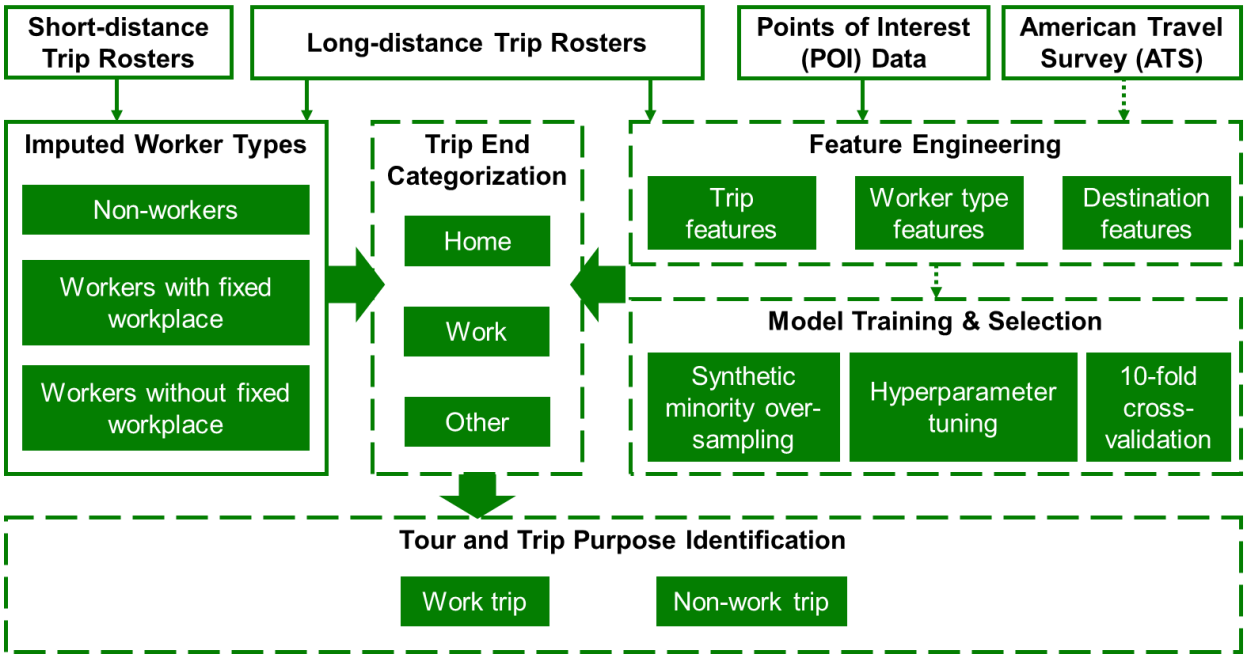


Figure 14. Flowchart of trip purpose imputation

3.5.1. Data Preparation

In this step, short-distance trips were categorized into three types of trip ends:

- If the trip end and the imputed home location of the corresponding traveler were at the same location, this trip end was labeled as "home"
- If the trip end and the imputed work location of the corresponding traveler (if the traveler has a work location) were at the same location, this trip end was labeled as "work"
- All the other trip ends were labeled as "other"

For the long-distance trips, a two-step model was implemented. First, the long-distance trip ends were matched with the Point of Interest (POI) data. If the traveler stayed for more than 2 hours within 656 ft (or 200 meters) of a POI of "Convention and Exhibition Center," between 8 a.m. – 8 p.m. on one day, this tour was labeled as a "Convention Center Staying," The duration of the stay was calculated as the time difference between the timestamp of the trip ends for the arrival at, and the departure from, the establishment.

For other long-distance trips, a machine learning model was applied. The feature selection for the machine learning model considered features that could be extracted from both the mobile device location data and the travel survey to ensure that the imputation model was applicable to the mobile device location data. The majority of the training dataset for long-distance trip purpose imputation in this project is the 1995 American Travel Survey (ATS), which is the most recent dataset of long-distance passenger travel information available at the national level. The selected features are listed in Table 3, which could be categorized into trip-related information, traveler-related information, and destination-related land use information.

Table 3. Features Selected for Long-distance Trip Purpose Imputation

Variable Category	Variable Name	Description
Trip-related information	#Trips/month	Number of long-distance trips per month
	Weekend trip	Indicates if the trip spanned a weekend (Saturday and Sunday)
	#Nights away	Number of nights away from home
	#Nights at destination	Number of nights at destination
	Principal transportation	Principal travel mode from origin to destination
	Great circle distance	Great circle distance from origin to destination
	#Stops to destination	Number of stops to destination
	#Side trips	Number of side trips
Traveler-related information	Worker	Whether the traveler is a worker
Destination-related information	Destination state	State of trip destination
	Destination region	Census Region of trip destination, including Northeast, Midwest, South, and West
	Destination census Division	Census Division of trip destination, including New England, Middle Atlantic, East North Central, West North Central, South Atlantic, East South Central, West South Central, Mountain, and Pacific
	Tourism	National park recreation visits by state
	GSP	Gross state product
	%Urban	Percentage of urban land use cover by state
	%Nature	Percentage of natural land use cover by state
	%Agriculture	Percentage of agriculture land use cover by state

3.5.2. Imputation Algorithm

3.5.2.1. Short-distance trip purposes

For short-distance trips, since the categories of the trip ends were identified in the data preparation step, the trip purpose was imputed based on the following rules: 1) trips with at least one trip end at the work location were identified as work trips; 2) all trips between two work trips were also identified as work trips; and 3) all other trips were identified as non-work trips.

3.5.2.2. Long-distance trip purposes

For long-distance trips, the trips were imputed based on the following rules: 1) all long-distance tours labeled with "Convention Center Staying" in the pre-processing step were identified as business tours. 2) the purpose of other long-distance tours was imputed by a machine learning model into one of two categories: business and non-business tours. All the trips in a business tour were considered work trips, and all the trips in a non-business tour were considered non-work trips.

3.6. Trip Distance Calculation

To produce reliable VMT statistics and trip distance distribution, it is important to develop an accurate trip distance estimation method. The prevailing method employed by commercial data providers is to either use the airline distance between origin and destination points, which drastically underestimates the actual trip distance on the transportation network (except air travel mode), or to use the shortest path algorithm assumptions and assignment of OD tables on a routable, multimodal transportation network. In this project, a scalable map matching and routing algorithm was incorporated to reconstruct the path of the driving and rail trips and then calculate their trip distances based on the observed travel routes. The detail of trip distance calculation for each specific mode is described below.

3.6.1. Map Matching and Routing

The UMD team developed and implemented a computationally efficient method for snapping sightings to routable transportation networks. A spatial index method, KD-Tree, was first used to find all the roads within 328 ft (or 100 meters) for each sighting. The next step was to construct the complete path between all the sightings snapped to the road networks using routing algorithms. For each sighting, the algorithm first compared its travel direction and the travel direction of its nearby roads within 328 ft. The closest candidate link with an absolute travel direction difference smaller than 30 degrees was selected as a valid match. Then, the path between the consecutive matched sightings was reconstructed by using the shortest path algorithm based on road length. In the meantime, reasonableness checks were also conducted during the routing process. For each pair of consecutive sightings snapped to the network, the routed distance was first calculated by adding the length of all the road segments routed between the two sightings. Then, two reasonableness checks were conducted (Newson and Krumm, 2009):

- If the routed distance was greater than the cumulative distance between the two observed snapped to the network by 1.24 miles or more, the route was considered invalid and in need of revision.
- The travel time on these links was calculated based on the timestamp difference of the two snapped sightings. With the routed distance and travel time, the average travel speed on these links were calculated. If the speed exceeded 112 mph (180 km/h), one of the two sightings was considered to be matched to the wrong link.

If either of these two violations was observed, an incremental approach was conducted by randomly removing one of the sightings, conducting the routing with the previous/next sighting snapped to the network, and examining the distance and travel speed until they did not violate the 1.24-mile threshold or the 112 mph threshold.

3.6.2. Mode-Specific Trip Distance Calculation

3.6.2.1. Vehicle travel

After implementing the aforementioned map matching and routing algorithm for vehicle trips, the complete path between all the sightings on the road network for each trip was constructed. Next, the trip distance was calculated as the sum of all segment lengths on the trip path.

3.6.2.2. Rail travel

Similar to the vehicle trips, all rail trip sighting points were snapped to the rail network and the trip distance was calculated after routing is implemented on the points. Considering that the rail network had significantly fewer links and limited route options compared to the road network, the map matching and routing had higher precision for the rail trips. The trip distance was similarly derived based on the length of the traversed segments for all unlinked rail trips.

To report the trip distance for linked rail trips that are comprised of multiple unlinked trips, the algorithm sums all the calculated distances of the respective unlinked trips and adds the gap distances between each consecutive unlinked trip pair.

3.6.2.3. Air travel

For air travel, the geodesic distance of the origin and destination of each trip was used as the trip distance. If the air trips had layovers, the geodesic distances between each flight segments were summed as the final trip distance.

3.6.2.4. Active transportation and ferry travel

For active transportation and ferry (ATF) travel modes, which mainly consist of walk, bike, and ferry, the map matching to the road network might not lead to an accurate reconstruction of the travel path, as pedestrians and bikers might decide to not follow the road networks for their trips. Therefore, for these trips, the method relied on the summation of the geodetic distances between all consecutive sightings for each trip.

In summary, Chapter 3 documented the methodological steps for trip data development. Key steps include trip identification, travel mode imputation, transit trip linking, worker type identification, trip purpose imputation, and trip distance calculation. As a result, the national trip roster was generated based on the national device and location data panel. The processed national trip roster then served as the input to the multi-level data expansion to form the national all trip roster (as shown in Figure 1 and elaborated in the next Chapter).

4. NATIONAL PASSENGER OD DATA DEVELOPMENT

This section describes the methodology for expanding the processed national passenger trip roster and aggregating the expanded national all trip roster into national passenger OD data products.

4.1. Data Expansion

Sighting data came from a non-probability sample of devices and did not cover the entire population of the U.S. Known biases associated with sighting data and the OD products derived from such data sources include but are not limited to the following:

- Different upstream data providers have access to different subsets of device owners.
- The owners of the devices in the sample do not represent the full population of the U.S. and are not equally representative of different socio-demographic groups.
- Data coverage may be different in urban and rural areas because of different mobile device penetration rates across the U.S.
- Not all movements of devices are necessarily observed. There is a higher probability of observing location records when the trip lasts longer, and the travel mode uses a transportation network with a more stable communication network.
- There are temporal biases in the location records of the observed devices due to different levels of mobile device usage during different hours of the day.

It is necessary to develop a proper data expansion procedure to generate population-representative statistics from a sample. For the NextGen NHTS OD program, a multi-level data expansion method was applied (device-level expansion and trip-level adjustment) to produce OD products that were representative of the entire U.S. population and its corresponding movements (Figure 15).

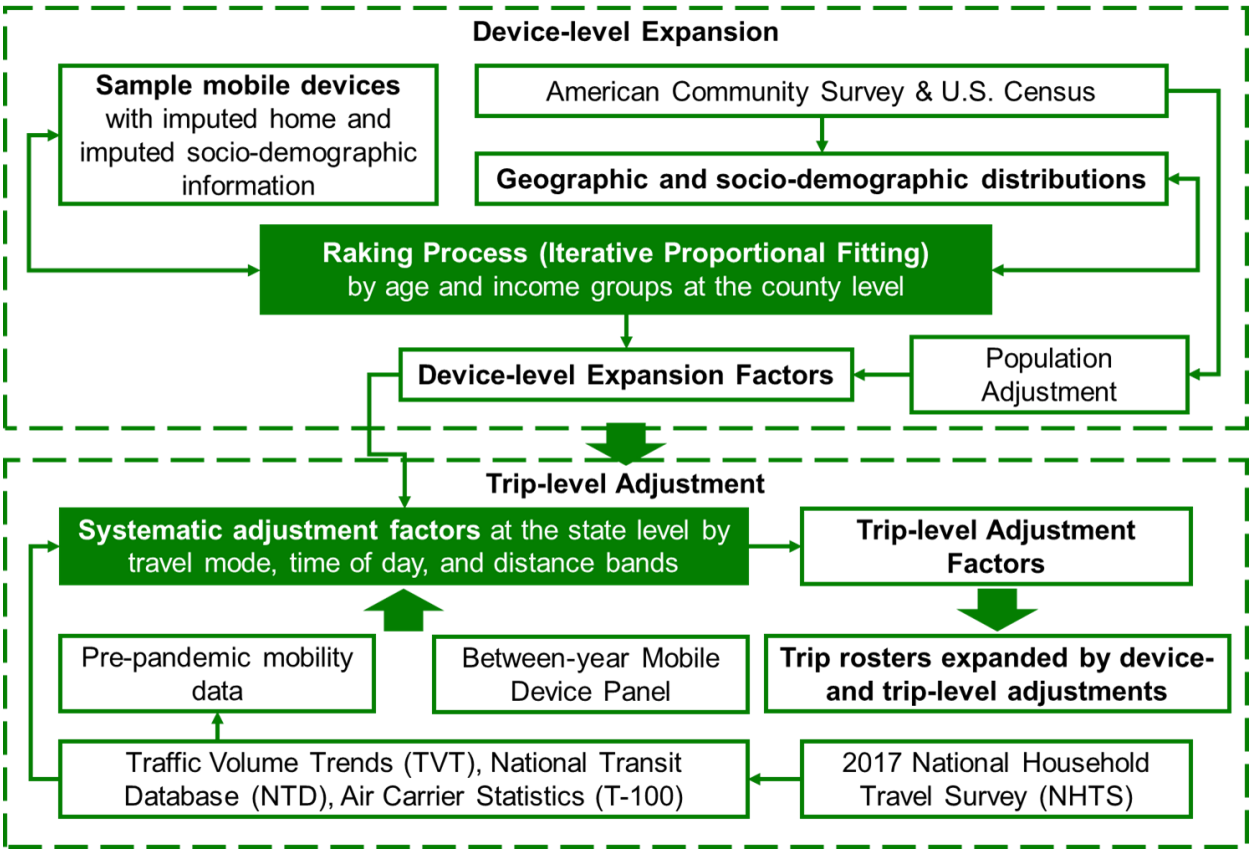


Figure 15. Flowchart of the multi-level data expansion

4.1.1. Device-Level Expansion

For device-level expansion, iterative proportional fitting (IPF), also known as the raking process, was used to expand the sample device estimates to the population-representative estimates. The first step was device selection. Only devices that passed the quality check and had home locations identified were considered in the device-level expansion. The monthly sampling rate (the number of devices per population) of such devices at the state level is shown in Figure 16. The overall effective sampling rate at the national level was 16.1%.

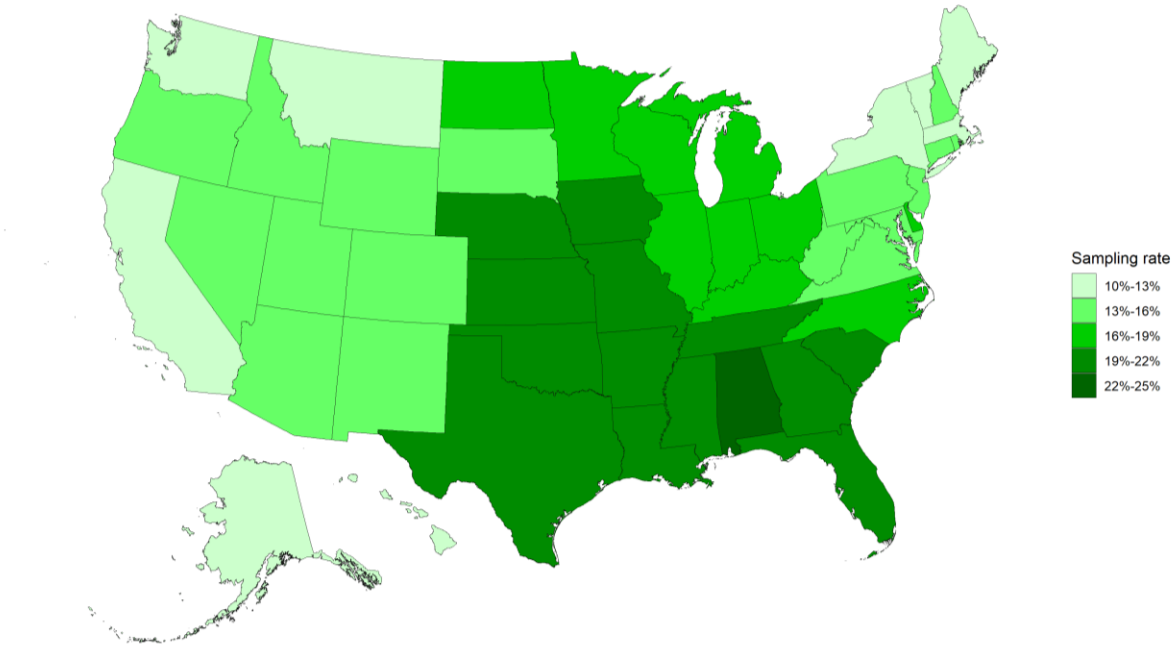


Figure 16. Effective sampling rate of the devices from the processed national trip roster employed in this project at the state level for 2020 OD data product

The second step was to address the device sample representativeness bias across different socio-demographic groups through socio-demographics imputation. A decision tree-based machine learning model was developed to impute device-level socio-demographic characteristics using trip-related information, traveler-related information, and home-related information derived and extracted from a large nationwide sighting dataset with true socio-demographic labels with over 400 thousand respondents. The model categorized device owners into five age groups—"less than 25 years old," "25-34 years old," "35-54 years old," "55-64 years old," and "65 years old and above"—and five income groups—"less than $25,000/year," "$25,000-$50,000/year," "$50,000-$75,000/year," "$75,000-$125,000/year," and "more than $125,000/year." All the cut points selected can be nested with the ACS categories, which were later used as control totals in the device-level expansion.

Figure 17 shows the framework of the device-level expansion based on the IPF method. UMD collected the latest 2015-2019 five-year county-level ACS data to obtain the control totals for the number of households, population by age and income groups. The age and income groups were further aggregated into 5 groups respectively, resulting in a total of 25 subcategories. The IPF method was then applied at the county level to generate a device-level expansion factor to match the control totals. If a certain county had zero observations in one of the 25 subcategories, the 25 subcategories for that county were aggregated into 9 subcategories to continue the IPF process. If there were still zero observations for one of the nine subcategories, the population-level expansion factors were applied, which were computed by dividing the county population by the number of devices residing in that county. Table 4 shows the subcategories for the 25-category and 9-category IPF, respectively.

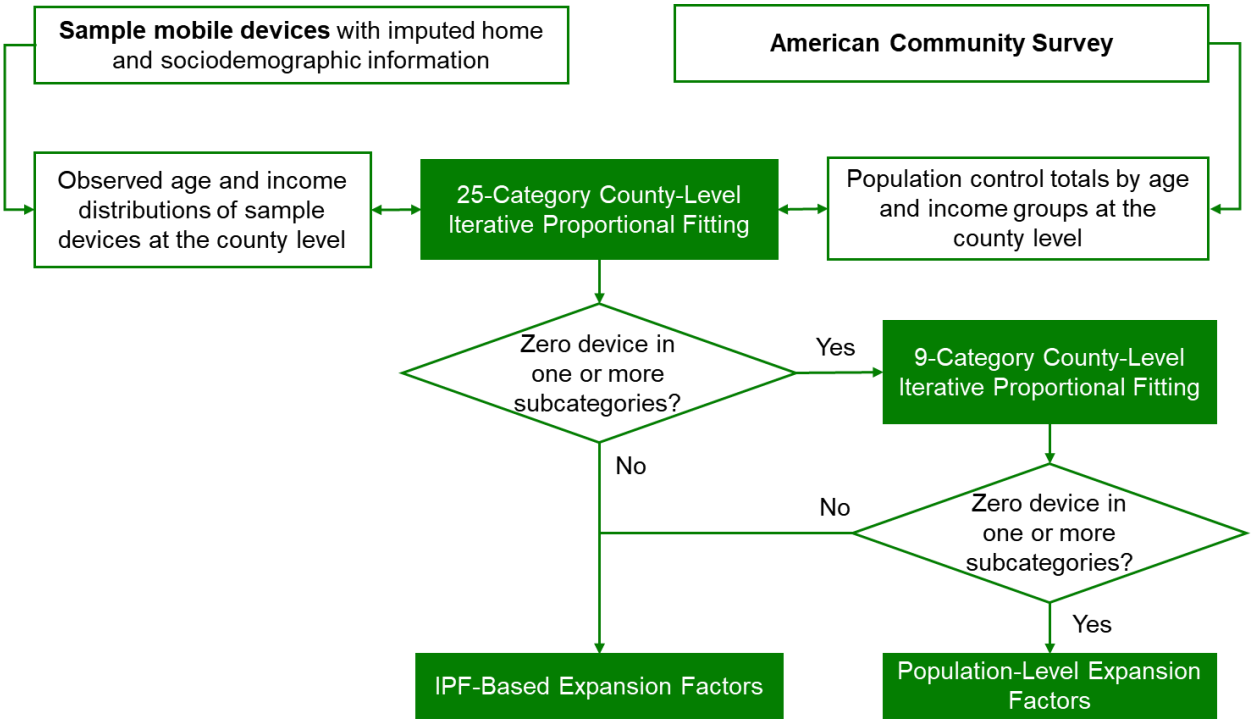


Figure 17. The framework for county-level iterative proportional fitting

Table 4. Categories Considered in the IPF

Initial Categories in the Twenty-Five-Category IPF	Aggregated Categories in the Nine-Category IPF
Less than 25 years old & less than $25,000/year Less than 25 years old & $25,000-$49,999/year 25-34 years old & less than $25,000/year 25-34 years old & $25,000-$49,999/year	Less than 35 years old & less than $50,000/year
Less than 25 years old & $50,000-$74,999/year Less than 25 years old & $75,000-$124,999/year 25-34 years old & $50,000-$74,999/year 25-34 years old & $75,000-$124,999/year	Less than 35 years old & $50,000-$124,999/year
Less than 25 years old & $125,000 and more/year 25-34 years old & $125,000 and more/year	Less than 35 years old & $125,000 and more/year
35-54 years old & less than $25,000/year 35-54 years old & $25,000-$49,999/year 55-64 years old & less than $25,000/year 55-64 years old & $25,000-$49,999/year	35-64 years old & less than $50,000/year
35-54 years old & $50,000-$74,999/year 35-54 years old & $75,000-$124,999/year 55-64 years old & $50,000-$74,999/year	35-64 years old & $50,000-$124,999/year

55-64 years old & $75,000-$124,999/year	
35-54 years old & $125,000 and more/year 55-64 years old & $125,000 and more/year	35-64 years old & $125,000 and more/year
65 years old and above & less than $25,000/year 65 years old and above & $25,000-$49,999/year	65 years old and above & less than $50,000/year
65 years old and above & $50,000-$74,999/year 65 years old and above & $75,000-$124,999/year	65 years old and above & $50,000-$124,999/year
65 years old and above & $125,000 and more/year	65 years old and above & $125,000 and more/year

After the expansion factors for all selected devices were estimated, a temporal adjustment factor of 1.0035 was calculated by dividing the 2020 U.S. population estimates from the Census by the 2019 estimates to account for the population growth. This temporal factor was then applied to all expansion factors to represent the 2020 U.S. population.

4.1.2. Trip-Level Adjustment

For the trips identified from mobile device data regarding time of day and trip distance, the major bias in trip estimates was two-fold: 1) the raw sightings of each device might not be complete during the observed time, and 2) the trip identification algorithms might introduce some systematic bias to the imputed trips. The trip-level adjustment was then necessary to address the inherited and systematic bias. The time-of-day bias is mainly related to the intuitive bias of mobile device (LBS) data collection that people's usage of smartphone apps is not evenly distributed throughout the day, thus impacting the sighting volume. Due to the differences in the detection of trips, the difference in trip distance distribution was widely discovered between passive data (mainly GPS survey) and survey data estimation. It was found that passive data yield higher trip rates, smaller trip distance and travel time, more driving trips, and lower non-motorized trips (Wang and Chen, 2018; Wang et al., 2019). Considering those biases and the impacts of the COVID pandemic, the team first calibrated the pre-pandemic mobility data by mode, departure time, and distance band using the ground truth estimates/trends from the 2017 NHTS, traffic volume trends (TVT) reports, national transit database (NTD), Air Carrier Statistics (T-100), etc. Then the national passenger trip rosters with device-level expansion factors were further adjusted based on the pre-pandemic mobility data and mode-wise travel trends observed from ground truth data (i.e., TVT, NTD, and T-100) and mobile device data panel. All the following adjustments were applied to the national passenger trip roster that was adjusted using the device-level expansion factors to derive the final multi-level expansion factors.

4.1.2.1. Air travel

The monthly T-100 domestic market data for all carriers served as the ground truth data source. For each month, the adjustment factors by origin and destination state and distance bands were developed based on the average daily mobile device air trip estimates from 14 benchmark days and the average daily reported trips from T-100. The average daily air trips from T-100 were calculated by dividing the reported monthly total departures/arrivals by the number of calendar

days in that month. The monthly state-level production and attraction were alternately adjusted until the errors arising from the comparison with the T-100 data no longer had significant decreases. The adjustment factors by origin and destination state and distance bands were then applied to the daily mobile device air trip estimates for that month to obtain the final expanded air trip estimates. The expanded monthly totals were summed to the expanded annual total.

4.1.2.2. Vehicle travel

The team employed the 2017 NHTS and the monthly VMT trend from the Traffic Volume Trends (TVT) reports as the ground-truth data source to first calibrate the pre-pandemic mobility data as a baseline. For each month, the team first generated the ground-truth vehicle trip estimates by inflating the 2017 NHTS vehicle trip estimates with the monthly VMT trend at the census division level. The inflation based on VMT assumed that the trip distance distribution did not change over time from the 2017 NHTS survey period to the pre-pandemic year. Then the adjustment factors by census division, departure time of day, and distance band were developed using the average daily mobile device vehicle trip estimates from 14 benchmark days and the average daily vehicle trip totals from the temporally-adjusted 2017 NHTS estimates. The adjustment factors by census division, departure time of day, and distance bands were applied to the daily mobile device vehicle trip estimates for that month to obtain the final expanded vehicle trip estimates as a baseline.

From the baseline, the vehicle trips were jointly adjusted using the VMT trends observed from the TVT reports and a mobile device data panel. The between-year mobile device panel was constructed from mobile device location data for each calendar month to adjust the travel behavior trends in terms of the distance band distribution at state level. For example, to capture the temporal changes in travel behaviors between January 2019 and January 2020, a mobile device panel was formed using mobile devices that provided high-quality mobility data in both months. The vehicle trip expansion framework was applied year by year, where the new year's products were adjusted based on the previous year's data. As new NextGen NHTS core data are released, UMD will use the newly released survey data to estimate a new baseline, which will then serve as the new foundation to adjust the following years' data products. The expanded monthly totals were summed to the expanded annual total.

4.1.2.3. Rail travel

The team employed the 2017 NHTS and the monthly rail trip trend from the National Transit Database (NTD) as the ground-truth data source to first calibrate the pre-pandemic mobility data as a baseline. For each month, the ground truth rail trip estimates were generated by inflating the 2017 NHTS rail trip estimates with the monthly NTD rail unlinked passenger trip (UPT) trend at the national level. Then the adjustment factors by census division, departure time of day, and distance band were developed using the average daily mobile device rail trip estimates from 14 benchmark days and the average daily rail trip totals from the temporally-adjusted 2017 NHTS estimates. The adjustment factors by census division, departure time of day, and distance band were applied to the daily mobile device rail trip estimates for that month to obtain the final expanded rail trip estimates as a baseline.

From the baseline, the rail trips were jointly adjusted using the UPT trends observed from NTD and a mobile device data panel (same as described in Section 4.1.2.2). The rail trip expansion framework was applied year by year, where the new year's products were adjusted based on the previous year's data. When a new baseline is developed based on the most recent release of the NextGen NHTS core survey data, it will then serve as the new foundation to adjust the following years' data products. The expanded monthly totals were summed to the expanded annual total.

4.1.2.4. Active transportation and ferry travel

The team employed the 2017 NHTS and the annual population trend from the U.S. Census as the ground-truth data source to first calibrate the pre-pandemic mobility data as a baseline. For the baseline year, the ground truth ATF trip estimates were generated by inflating the 2017 NHTS ATF trip estimates with the annual population trend at the national level. Then the adjustment factors by census division, departure time of day, and distance band were developed using the average daily mobile device ATF trip estimates from 14 benchmark days in each month and the average daily ATF trip totals from the temporally-adjusted 2017 NHTS estimates. The adjustment factors by census division, departure time of day, and distance band were then applied to the daily mobile device ATF trip estimates for the entirety of 2020 to obtain the final expanded ATF trip estimates as a baseline.

From the baseline, the ATF trips were jointly adjusted using the population trends observed from the U.S. Census and a mobile device data panel (same as described in Section 4.1.2.2). The ATF trip expansion framework was applied year by year, where the new year's products were adjusted based on the previous year's data. When a new baseline is developed based on the more recent release of the NextGen NHTS core survey data, it will then serve as the new foundation to adjust the following years' data products. The expanded daily numbers of trips were summed to the expanded annual total.

4.1.3. Trip Distance Distribution Comparison

Figure 18 compares the trip distance distribution for unexpanded and expanded passenger trips. The expansion process increased the share of trips shorter than 10 miles, decreased the share of trips between 10 and 100 miles, and barely influenced the share of trips longer than 100 miles. The distribution trend among different distance bins remained relatively unchanged, which implied that the data expansion process did not distort the travel trends observed from the sighting data.

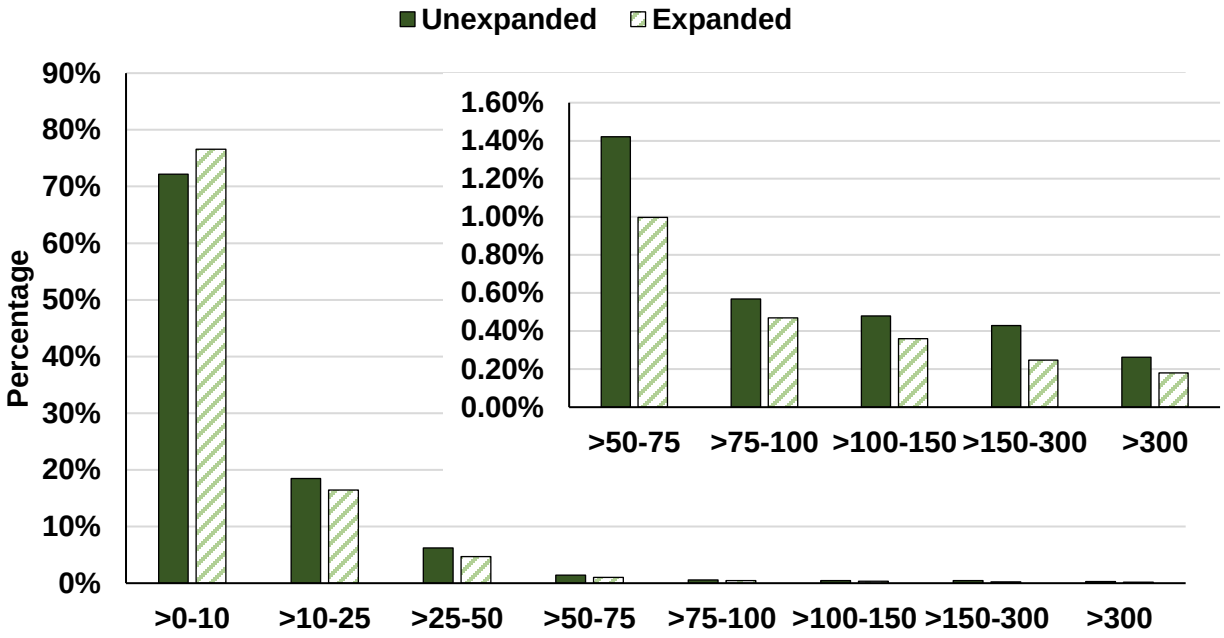


Figure 18. A comparison of distance distribution between unexpanded and expanded trips

4.2. Aggregating Trip Roster into a National Passenger OD Product

The team used the entire national passenger trip roster and the multi-level data expansion to develop the national passenger OD product. Figure 19 illustrates the resulting national annual average daily passenger trip production rates for 2020. In addition to trip rates, other critical analytics such as trip distance distribution, passenger mode share, and trip purpose were generated as product features.

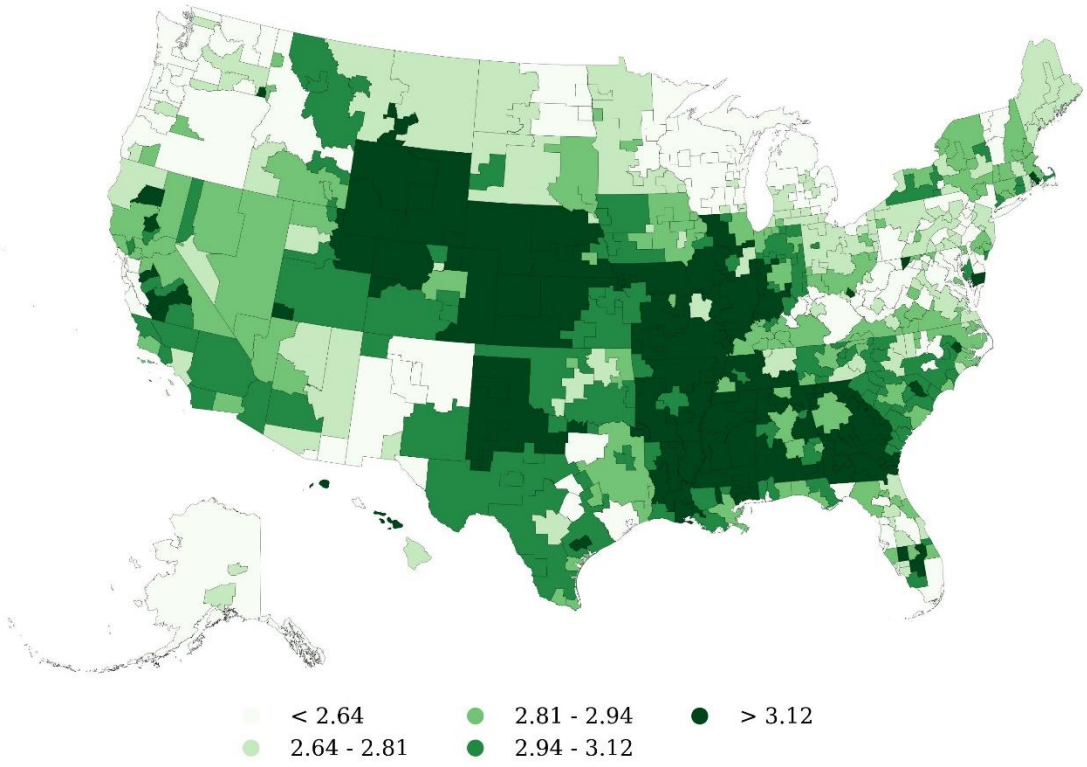


Figure 19. National passenger trip production rate heatmap (2020 annual average)

To protect the privacy of devices in the national device and location data panel, all travel information of OD pairs with total annual number of trips less than 30 were reset to 0 trips and a "flag" column was added to the files with the value as "1" to denote masking. In this process, 86,031 OD pairs were impacted and 977,720 total annual trips were removed from the 2020 annual file (0.003% of the total annual trips).

The first version of the 2020 national passenger OD file included balanced OD pair, which was removed from the second version of the 2020 product. This modification was made to accommodate the introduction of monthly files (in addition to the annual product) with the 2021 national passenger OD product. Investigations on the imbalance of OD flows showed a 0.05% imbalance flow at the OD pair level based on the total volume of trips in the 2020 OD files.

5. VALIDATION PLAN

The UMD team developed a rigorous plan to validate the proposed algorithms and the final national OD data products at an aggregate level to ensure product quality and transparency. The aggregate-level validation plan is described in this chapter. Product validation targets were established using the core NHTS survey, the NTD, the DB1B data, the T-100 data, Highway Performance Monitoring System (HPMS), and other available datasets.

5.1. Validation of the National Passenger OD Data Product

The team conducted both an internal and an external quality assurance and quality control (QAQC) of the national passenger OD data product. Both the internal and external QAQC followed a similar procedure assessing the key elements of the national product, as outlined in this section.

Overall, the team compared the annual average daily trip rates computed from the national passenger OD data trip totals and the ACS population data with the 2017 NHTS daily trip rates at the census division level (as 2017 NHTS was the most recent dataset available). Both the passenger OD and the 2017 NHTS trip rates had similar spatial trends across different census divisions. Passenger trips were also validated by travel mode as described in Sections 5.1.1, 5.1.2, and 5.1.3.

5.1.1. National Vehicle Passenger Trips

For the vehicle mode, vehicle miles traveled (VMT) was selected as the metric to be evaluated. As the national passenger OD data report person trips, the vehicle occupancy was first estimated for each person trip so that the person miles traveled (PMT) could be converted to VMT. The UMD team compared the annual average daily VMT per person computed from the national passenger OD data and the ACS population data with that computed from the 2020 TVT VMT data and the ACS population data since the 2020 HPMS data were not available at the point in time when validation was conducted. Both the passenger OD and TVT VMT per person estimates had similar spatial trends across different census divisions.

5.1.2. National Air Passenger Trips

The UMD team leveraged the 2020 DB1B data and the T-100 data as the calibration and validation data sources for air trip validaiton. DB1B is a 10% sample of all itineraries flown on all domestic certificated route carriers and intra-Alaska carriers. It is reported quarterly and has three data tables: Ticket, Market, and Coupon. The Ticket data report the entire itineraries, the Market data report the layovers, and the Coupon data report all trip legs. T-100 data provide monthly traffic for each operating carrier and its corresponding market and represent a full enumeration of the entire U.S. population. It has two data tables for domestic flights: Domestic Market and Domestic Segment. The Domestic Market data report trips by flight number, which may include interim stops. The Domestic Segment data report all non-stop flights like DB1B Coupon data.

Since the national passenger OD data reports air trips without layovers, the UMD team calibrated the air mode imputation parameters with the DB1B Market data (see Section 3.2.1) and validated the air trip estimates with the T-100 Domestic Market data. The total national air passenger trips from the OD data were compared with the T-100 data. The annual total percentage discrepancy was -0.54%, which met the contract requirement that the discrepancy should be within +/- 10%.

5.1.3. National Rail Passenger Trips

The UMD team leveraged the 2020 NTD data as the calibration and validation data sources for the rail trips. Since NTD data report unlinked passenger trips (UPTs) by transit agencies, the average number of transfers was calculated from the 2017 NHTS survey data to convert the linked rail trip estimates from the OD data to unlinked ones for a consistent comparison with the NTD UPT estimates. The annual absolute percentage difference between the two data was -2.94%, which met the contract requirement that the discrepancy should be within +/- 10%.

5.1.4. Additional Quality Control

In addition to the aforementioned validation process, the UMD team produced the following tabulations of national passenger OD data for future comparison with the core NHTS survey estimates and other historical NHTS data: (1) modal share percentages data; (2) trip purpose share percentages data; (3) trip length distribution; (4) trip length distribution by mode; (5) trip length distribution by trip purpose; (6) modal share by trip purpose; (7) trip length distribution by mode and trip purpose.

5.2. Reasonableness Check

In addition to the aforementioned validation plan, the UMD team examined the national passenger OD data to ensure that the data had no extreme or unreasonable values in any geography and were logically reasonable.

The team confirmed that the national passenger OD data did not have: (1) “ATF” mode trips longer than 100 miles, i.e., extremely long non-motorized trips; (2) trips between Alaska and the contiguous United States (48 adjoining U.S. states plus the District of Columbia) by ground transportation modes, including “vehicle”, “rail”, and “ATF” modes; (3) trips between the continental United States and Hawaii by ground transportation modes, including “vehicle”, “rail”, and “ATF” modes; (5) trips by “vehicle” and “rail” modes between the three zones in Hawaii; (6) extremely short air trips shorter than 75 miles.

6. REFERENCES

Alexander, L., Jiang, S., Murga, M., & González, M. C. (2015). Origin–destination trips by purpose and time of day inferred from mobile phone data. Transportation research part c: emerging technologies, 58, 240-250.

Batini, C. & Scannapieco, M. (2006). Data Quality: Concepts, Methodologies and Techniques. Springer Verlag.

Bohte, W. & Maat, K. (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, 17(3), 285-297.

Cappiello, C. et al. (2003). Data quality assurance in cooperative information systems: a multi-dimension certificate. In proceedings of the ICDT international workshop on data quality in cooperative information systems.

Chen, C., Bian, L., & Ma, J. (2014). From sightings to activity locations: how well can we guess the locations visited from mobile phone sightings. Transportation Research Part C: Emerging Technologies, 46(10), 326-337.

Chen, C., Ma, J., Susilo, Y., Liu, Y., & Wang, M. (2016). The promises of big data and small data for travel behavior (aka human mobility) analysis. Transportation Research Part C: Emerging Technologies, 68, 285-299.

Clark, H. M. (2017). Who Rides Public Transportation. American Public Transportation Association. https://www.apta.com/research-technical-resources/research-reports/who-rides-public-transportation/.

Gini, C.: Variabilità e Mutuabilità. (1912). Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche. C. Cuppini, Bologna.

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

Huang, L., Li, Q., & Yue, Y. (2010). Activity identification from GPS trajectories using spatial temporal POIs' attractiveness. In Proceedings of the 2nd ACM SIGSPATIAL International Workshop on location based social networks, 27-30, ACM.

Newson, P. & Krumm, J. (2009, November). Hidden Markov map matching through noise and sparseness. In Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems (pp. 336-343).

Pan, Y., Sun, Q., Yang, M., Darzi, A., Zhao, G., Kabiri, A., Xiong, C., & Zhang, L. (2023). Residency and worker status identification based on mobile device location data. Transportation Research Part C: Emerging Technologies, 146, 103956.

Phithakkitnukoon, S., Horanont, T., Di Lorenzo, G., Shibasaki, R., & Ratti, C. (2010). Activity-aware map: Identifying human daily activity pattern using mobile phone data. In International Workshop on Human Behavior Understanding, 14-25, Springer, Berlin, Heidelberg.

Stenneth, L., Wolfson, O., Yu, P. S., & Xu, B. (2011, November). Transportation mode detection using mobile phones and GIS information. In Proceedings of the 19th ACM SIGSPATIAL international conference on advances in geographic information systems (pp. 54-63).

U.S. Bureau of Labor Statistics. (2020). 2017, 2018, and 2019 American Time Use Survey (ATUS). https://www.bls.gov/tus/.
U.S. Bureau of Labor Statistics. (2020). 2020 Occupational Employment and Wage Statistics (OEWS). https://www.bls.gov/oes/.
U.S. Bureau of Labor Statistics. (2020). 2018 Standard Occupational Classification (SOC) System. https://www.bls.gov/soc/2018/major_groups.htm.
U.S. Census Bureau. (2020). American Community Survey (ACS). https://www.census.gov/programs-surveys/acs.
U.S. Census Bureau. (2020). Longitudinal Employer-Household Dynamics (LEHD) Origin-Destination Employment Statistics (LODES). https://lehd.ces.census.gov/data/#lodes.
U.S. Department of Homeland Security. Homeland Infrastructure Foundation-Level Data (HIFLD). https://hifld-geoplatform.opendata.arcgis.com/.
U.S. Department of Transportation, Bureau of Transportation Statistics. (2021). Air Carrier Statistics (Form 41 Traffic)- All Carriers. 2020 T-100 Domestic Market (All Carriers). https://www.transtats.bts.gov/DatabaseInfo.asp?QO_VQ=EEE&Yv0x=D.
U.S. Department of Transportation, Bureau of Transportation Statistics. (2021). Airline Origin and Destination Survey (DB1B). 2020 DB1BMarket. https://www.transtats.bts.gov/Fields.asp?gnoyr_VQ=FHK.
U.S. Department of Transportation, Bureau of Transportation Statistics. (2021). 1995 American Travel Survey (ATS). https://www.bts.gov/browse-statistical-products-and-data/surveys/american-travel-survey.
U.S. Department of Transportation, Bureau of Transportation Statistics. 2020 Amtrak Ridership Data.
U.S. Department of Transportation, Federal Highway Administration, 2017 National Household Travel Survey. 2017 NHTS Data User Guide. https://nhts.ornl.gov/assets/2017UsersGuide.pdf.
U.S. Department of Transportation, Federal Highway Administration, Office of Highway Policy Information. Highway Performance Monitoring System (HPMS). https://www.fhwa.dot.gov/policyinformation/hpms.cfm.
U.S. Department of Transportation, Federal Highway Administration, Office of Highway Policy Information. 2020 Travel Volume Trends. https://www.fhwa.dot.gov/policyinformation/travel_monitoring/tvt.cfm.
U.S. Department of Transportation, Federal Motor Carrier Safety Administration (FMCSA). (2020). Summary of Hours of Service (HOS) Regulations. https://www.fmcsa.dot.gov/regulations/hours-service/summary-hours-service-regulations.
U.S. Department of Transportation, Federal Transit Administration, the National Transit Database (NTD). Monthly Module Adjusted Data Release. https://www.transit.dot.gov/ntd.
Wang, F., & Chen, C. (2018). On data processing required to derive mobility patterns from passively-generated mobile phone data. Transportation Research Part C: Emerging Technologies, 87, 58-74.

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

Xie, K., Deng, K., & Zhou, X. (2009). From trajectories to activities: a spatio-temporal join approach. In Proceedings of the 2009 International Workshop on Location Based Social Networks, 25-32, ACM.

Yang, M., Pan, Y., Darzi, A., Ghader, S., Xiong, C., & Zhang, L. (2021). A data-driven travel mode share estimation framework based on mobile device location data. Transportation, 1-45.

Zhang, L., Ghader, S., Darzi, A., Pan, Y., Yang, M., Sun, Q., Kabiri, A., & Zhao, G. (2020). Data analytics and modeling methods for tracking and predicting origin-destination travel trends based on mobile device data. Federal Highway Administration Exploratory Advanced Research Program.

GLOSSARY

Active Local Hours	Mean, 25th, 50th and 75th percentile of the average daily number of local hours observed for RAUs.
Activity Location Identification	The methodology to identify the most significant locations for each device from a set of activity locations from the sighting data.
Business Tour	A tour with business activities as the primary trip purpose.
Daily Active Users (DAUs)	The number of devices with at least one sighting on a specific day for a specific month.
Daily Coverage	Variance in the total sighting volume by day of month for all RAUs, measured by a Gini coefficient between 0 and 1, with 0 indicating equal total number of sightings from each day in one month and 1 indicating all sightings are from one day.
Data Frequency	Mean, 25th, 50th and 75th percentile of average daily number of sightings by RAU devices.
Data Oscillation	Abnormal movements with unreasonable distance and time combinations between sightings.
Data Preprocessing	Data cleaning steps including removal of sightings with invalid data entries, removal of duplicate sightings, removal of data oscillations, etc.
Device Deduplication	The methodology to deduplicate the devices potentially owned by the same individual from the sighting data.
Device-Level Expansion	The methodology to expand the sample devices to represent the entire population of the U.S.
Device Representativeness	Variance in average daily number of sightings among RAU devices, measured by a Gini coefficient between 0 and 1, with 0 indicating equal sighting frequency and 1 indicating distinct sighting frequency for all RAUs.
Fixed Workplace Location Identification	The methodology to identify the fixed workplace location of a device from the sighting data if it exists.
Geographical Representativeness (by Device)	Variance of population coverage among different counties, measured by a Gini coefficient[2] between 0 and 1, with 0 indicating equal sampling rate in all zones and 1 indicating that all RAUs are from a single zone.

[2] Gini coefficient (Gini, 1912) is a statistical measure of the equality of a given data. It can be calculated by the ratio of the area above the Lorenz curve to the summation of the area above and the area below the Lorenz curve. The Lorenz curve is a graph showing the distribution of the given data.

Geographical Representativeness (by Sighting)	Variance of sighting volume divided by county-level population, measured by a Gini coefficient between 0 and 1, with 0 indicating equal sighting volume per person in all zones and 1 indicating that all sightings are from a single zone.
Geohash	A public domain geocode system that encodes a geographic location into a short string of letters and digits. There are twelve levels of geohash zones, which differ in zone size, length of the zone name, etc.
Home Location Identification	The methodology to identify the home location of a device from the sighting data.
Hourly Coverage	Variance in the average sighting volume by hour of day for all RAUs, measured by a Gini coefficient between 0 and 1, with 0 indicating equal average number of sightings from the 24 hours and 1 indicating all sightings are from one hour.
Linked Trip	For short-distance trips, a sequence of unlinked trips made between a series of locations joined together based on the primary travel mode (i.e., transit modes). For long-distance trips, a linked trip can be a sequence of non-air trips made between primary stops, a sequence of air trips between primary stops, or the access and egress trips to the air trips between primary stops. For all other cases, the unlinked trips were directly treated as linked trips.
Location Accuracy	Mean, 25^{th}, 50^{th} and 75^{th} percentile of the positioning accuracy of RAU devices. Positioning accuracy is defined as the maximum distance between a device's recorded location and its actual location at 95% confidence level.
Location Recording Interval	The time duration between the consecutive sightings.
Location Sighting	A location sighting is generated when a mobile application updates the device's location with the most accurate sources based on the existing location sensors such as Wi-Fi, Bluetooth, cellular tower, and Global Positioning System (GPS). It usually records an anonymized device identifier (ID), latitude and longitude coordinates, time stamps, positioning accuracy, etc.
Long-Distance Tour	Tours in which a device is observed equal to or more than 50 miles away from the home location.
Long-Distance Trip	Trips within long-distance tours.
Long-Distance Trip Purpose	The trip purpose for long-distance trips.

Monthly Active Users (MAUs)	The number of devices with at least one sighting for a specific month.
Multi-Level Data Expansion	The methodology composed of device-level expansion and trip-level adjustment.
National Device and Location Data Panel	The cleaned device and location data panel is developed from the raw sighting data panel through data preprocessing, quality assessment, home and fixed workplace location identification, and device deduplication and sighting data integration, which is then used for identifying trip-level information.
National All-Trip Roster	The trip roster based on the national device and location data panel.
National Passenger OD Data Product	The expanded number of trips for each OD pair representing the population travel within and between the zones by trip distance, trip purpose, travel modes, etc., based on the national all-trip roster.
Non-business Tour	A tour with non-business activities as the primary trip purpose, such as personal recreation.
Passively Collected Location Data/Mobile Device Location Data	Location sighting data generated by mobile devices, e.g., cell phones and tablets, from various positioning technologies such as cellular networks, GPS, and location-based services (LBS).
Primary Destination	The farthest primary stop that is located at least 50 miles away from the home location.
Primary Stop	A secondary stop where the device stays for a significant amount of time and/or from which the device makes local trips on a long-distance tour.
Professional Driver Identification	Drivers that regularly have driving trips with long trip durations.
Raw Sighting Data Panel	The raw sighting data panel consists of sighting data aggregated from multiple data vendors that consist of more than 270,000,000 monthly active users and that represent the movements across the nation.
Regularly Active Users (RAUs)	The number of devices with at least seven days of more than ten daily sightings for a specific month.
Scalable Map Matching and Routing	The methodology to snap the sighting data to the road network and estimate the path using a routing algorithm.
Secondary Stop	A place where the device stays for more than 30 minutes on a long-distance tour.
Segment-Based Approach	The methodology to impute the travel mode for a segment of the trip.

Short-Distance Tour	Tours in which a device is observed less than 50 miles away from the home location during the whole period of the tour.
Short-Distance Trip	Trips within short-distance tours.
Short-Distance Trip Purpose	The trip purpose for short-distance trips.
Sighting Data Integration	The methodology to integrate sighting data from different devices owned by the same individual and from different data providers.
Socio-Demographic Imputation	The methodology to impute socio-demographic information for each device.
Subtour	A segment of a long-distance tour that falls between two primary stops.
Temporal Adjustment Factor	The factor to account for the population growth from 2019 to 2020.
Temporal Consistency	The average number of observed days for RAUs for a specific month.
Temporal Similarity Ratio	A ratio to measure the similarity between unique hours when a device is observed at workplace candidates and an identified home location.
Tour/Home-Based Tour	A sequence of unlinked trips between the departure from and the arrival at one's identified home location.
Tour-Based Method	The methodology that first identifies the tours and enables one to consider trip linking and differentiate between linked and unlinked trips.
Tour and Trip Identification	The methodology to identify tours and trips from the sighting data.
Travel Mode Imputation	The methodology to impute the travel mode for unlinked trips from the sighting data.
Trip	Unlinked and linked trips.
Trip-Based Approach	The methodology to impute the travel mode for the entire trip.
Trip Distance Calculation	The methodology to estimate the trip distance for each identified trip using the sighting data and road network.
Trip-Level Adjustment	The methodology to expand the sample trips to represent the travel of the entire U.S. population based on control totals from external ground truth data.
Trip Linking	The methodology to merge the unlinked trips into linked trips.

Trip Purpose Imputation	The methodology to impute trip purpose for linked trips from the sighting data.
Unlinked Trip	The basic unit of analysis for trips.
Worker Type Identification	The methodology to identify the professional drivers and other workers without fixed workplaces.
Workers without Fixed Workplace	Workers—such as cleaning and pest control workers—that do not have a fixed workplace.