Next Generation National Household Travel Survey National Origin Destination Data

Passenger Origin-Destination Data Methodology Documentation
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>4. Title and Subtitle</th>
<th>5. Report Date</th>
<th>6. Performing Organization Code:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Next Generation National Household Travel Survey National Origin Destination Data Passenger Origin-Destination Data Methodology Documentation</td>
<td>October 2021</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Lei Zhang, Aref Darzi, Yixuan Pan, Mofeng Yang, Qianqian Sun, Aliakbar Kabiri, Guangchen Zhao, Chenfeng Xiong</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>9. Performing Organization Name and Address</th>
<th>10. Work Unit No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>University of Maryland Office of Research Administration 3112 Lee Building College Park, MD 20742-5141</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>11. Contract or Grant No.</th>
<th>12. Sponsoring Organization Name and Address</th>
<th>13. Type of Report and Period Covered</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>15. Supplementary Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Federal Highway Administration’s Task Monitors for this project were Daniel Jenkins, P.E. and Dr. Patrick Zhang, P.E.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>16. Abstract</th>
</tr>
</thead>
<tbody>
<tr>
<td>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.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>17. Key Words</th>
<th>18. Distribution Statement:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Origin destination passenger data, passively collected location data, travel behavior, emerging technologies, data privacy, National Household Travel Survey (NHTS), NextGen NHTS</td>
<td>Internal draft. NOT for public distribution.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Unclassified</td>
<td>Unclassified</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

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

Federal Highway Administration (FHWA) has launched the Next Generation National Household Travel Survey (NextGen NHTS) program, aimed at establishing a more continuous national travel monitoring program. As part of the NextGen NHTS program, this project produces 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. 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 weighted based on control totals at various levels. A national weighted 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 weighting 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 will be published by FHWA in the public domain, along with the associated source codes for the computation algorithms used in the development of these products.
TABLE OF CONTENTS

EXECUTIVE SUMMARY ................................................................................................................................. I
TABLE OF CONTENTS ................................................................................................................................. II
LIST OF FIGURES ....................................................................................................................................... IV
LIST OF TABLES .......................................................................................................................................... V

1. OVERVIEW OF THE TECHNICAL APPROACH AND METHODOLOGY ....................................................... 1

2. RAW SIGHTING DATA ASSEMBLY, PREPROCESSING, AND QUALITY EVALUATION ............................ 3
   2.1. Data Preprocessing and Quality Metrics .............................................................................................. 3
       2.1.1. Data Preprocessing ...................................................................................................................... 5
       2.1.2. Data Quality Metrics .................................................................................................................. 7
   2.2. The Identification of Home and Fixed Workplace ................................................................................. 9
       2.2.1. Home Location Identification ..................................................................................................... 11
       2.2.2. Fixed Workplace Location Identification .................................................................................... 12
   2.3. Device Deduplication and Sighting Data Integration .......................................................................... 13

3. NATIONAL PASSENGER TRIP DATA DEVELOPMENT ........................................................................... 16
   3.1. Tour and Trip Identification .................................................................................................................. 16
       3.1.1. Home-Based Tour Identification .................................................................................................. 16
       3.1.2. Trip Identification for Short-Distance Tours ................................................................................ 17
       3.1.3. Trip Identification for Long-Distance Tours ............................................................................... 19
           3.1.3.1. Stop and primary destination identification ........................................................................ 19
           3.1.3.2. Subtour identification ......................................................................................................... 19
           3.1.3.3. Trip identification ............................................................................................................... 19
   3.2. Travel Mode Imputation ....................................................................................................................... 21
       3.2.1. Air Travel Mode Imputation ........................................................................................................ 23
       3.2.2. Land and Water Transportation Travel Mode Imputation .......................................................... 23
           3.2.2.1. Feature engineering .............................................................................................................. 23
           3.2.2.2. Random forest model and its accuracy ................................................................................ 24
   3.3. Merging Unlinked Trip Segments into Trips ....................................................................................... 25
   3.4. Worker Type Identification ................................................................................................................. 26
       3.4.1. Professional Driver Identification ................................................................................................ 26
       3.4.2. Other Workers without Fixed Workplaces .................................................................................... 28
   3.5. Trip Purpose Imputation ..................................................................................................................... 28
       3.5.1. Data Preparation ......................................................................................................................... 29
3.5.2. Imputation Algorithm ............................................................................................................... 30
  3.5.2.1. Short-distance trip purposes .............................................................................................. 30
  3.5.2.2. Long-distance trip purposes ............................................................................................. 30

3.6. Trip Distance Calculation ........................................................................................................ 31
  3.6.1. Map Matching and Routing ................................................................................................ 31
  3.6.2. Mode-Specific Trip Distance Calculation ............................................................................. 31
    3.6.2.1. Vehicle travel .................................................................................................................. 31
    3.6.2.2. Rail travel ....................................................................................................................... 32
    3.6.2.3. Air travel ......................................................................................................................... 32
    3.6.2.4. Active transportation/ferries travel ................................................................................ 32
  3.6.3. Trip Distance Calculation for Linked Trips ............................................................................. 32

4. NATIONAL PASSENGER OD DATA DEVELOPMENT ................................................................... 33
  4.1. Weighting and Data Expansion ................................................................................................. 33
    4.1.1. Device-Level Weighting ..................................................................................................... 34
    4.1.2. Trip-Level Weighting ......................................................................................................... 37
      4.1.2.1. Air travel ....................................................................................................................... 38
      4.1.2.2. Vehicle travel ............................................................................................................... 38
      4.1.2.3. Rail travel .................................................................................................................... 38
      4.1.2.4. Active transportation/ferries travel ............................................................................. 39
    4.1.3. Trip Distance Distribution Comparison ................................................................................. 39
  4.2. Aggregating Trip Roster into a National Passenger OD Product .............................................. 40

5. VALIDATION PLAN ......................................................................................................................... 41
  5.1. Validation of the National Passenger OD Data Product ............................................................ 41
    5.1.1. National Vehicle Passenger Trips ....................................................................................... 41
    5.1.2. National Air Passenger Trips .............................................................................................. 41
    5.1.3. National Rail Passenger Trips ............................................................................................ 42
    5.1.4. Additional Quality Control .................................................................................................. 42
  5.2. Reasonableness Check ............................................................................................................. 42

6. REFERENCES ................................................................................................................................. 43

GLOSSARY ............................................................................................................................................. 46
LIST OF FIGURES

Figure 1. Origin-destination data production flow chart for the Next Generation National Household Travel Survey (NextGen NHTS) OD Data Program .................................................................................................................. 1
Figure 2. 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 OD data products .................................................................................................................. 4
Figure 3. Schema of the data quality evaluation and data preprocessing ........................................................................ 5
Figure 4. Procedure for removing data oscillations .................................................................................................................. 6
Figure 5. Two scenarios of data oscillations considered by Heuristic 2 ................................................................................. 7
Figure 6. The framework for home, fixed work locations, and worker type imputation ...................................................... 11
Figure 7. Flowchart of device deduplication and sighting data integration ............................................................................ 14
Figure 8. Tour identification and trip chaining demonstration ................................................................................................ 16
Figure 9. Recursive algorithm for trip identification for short-distance tours ........................................................................ 18
Figure 10. Recursive algorithm for trip identification for long-distance tours ............................................................. 20
Figure 11. Flowchart of travel mode imputation .......................................................................................................................... 22
Figure 12. Flowchart of merging unlinked trip segments into trips .................................................................................. 25
Figure 13. Flowchart of removing professional driver trips .................................................................................................. 27
Figure 14. Flowchart of trip purpose imputation ......................................................................................................................... 29
Figure 15. Flowchart of the multi-level weighting ...................................................................................................................... 34
Figure 16. Sampling rate of the devices from the unweighted national trip roster employed in this project at the state level for 2020 OD data products ............................................................................. 35
Figure 17. The framework for county-level iterative proportional fitting ............................................................................ 36
Figure 18. A comparison of distance distribution between unweighted and weighted trips .................................................. 39
Figure 19. National passenger trip production rate heatmap (2020 annual average) .......................................................... 40
LIST OF TABLES

Table 1. Literature Review on Travel Mode Imputation Methods ........................................................... 21
Table 2. Features for Detecting Land and Water Transportation Travel Mode ....................................... 24
Table 3. Features Selected for Long-distance Trip Purpose Imputation .................................................. 30
Table 4. Categories Considered in the IPF ................................................................................................ 37
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 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 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 national passenger OD data products.

The entire national all-trip roster was used by the UMD team to develop national passenger OD products. Trips were weighted based on population and employment data, imputed socio-demographics, and a multi-level weighing method that employed weights 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), the 2020 Amtrak ridership data, 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 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.
Figure 2. 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 OD data products
Various dimensions of assessing data quality, such as consistency, accuracy, completeness, and timeliness, have been 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 assesses the raw sighting data quality from the four dimensions, addresses the quality issues through data preprocessing, and evaluates 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](image)

### 2.1.1. Data Preprocessing

Raw sighting data is 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 are 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) are utilized in the current and following data processing steps. The simplified trajectories are utilized for detecting oscillations.

If a device is 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 are treated as true visits and form a “stable community.” Based on the identified true visits, other locations are investigated to check oscillations. All level-7 geohash zones involved in a stable community are determined to be stable level-7 geohash zones. Communities and level-7 geohash zones are used to remove oscillations in different cases.

Two heuristic rules are designed to remove oscillations:

**Figure 4. Procedure for removing data oscillations**
• Heuristic 1 at the geohash zone level: if a device leaves a stable level-7 geohash zone and returns to the same zone within 30 seconds, the sightings out of the stable zone during the 30 seconds are determined to be oscillations and are removed. In addition, if a device moves 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 are 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 moves to a faraway community C2 at high speed, the corresponding sightings in C2 are removed; (b) if the device moves 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 are considered oscillations and their corresponding sightings are removed. Specifically, the nearby and faraway communities are relative positions decided by the spatial-temporal criteria from Heuristic 1. The criteria utilize 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 happens, such as the continuously unstable communities, C1, C2, …, and C5, shown in Figure 5 (b), the dwell time of each group of communities is used for reserving one group of communities as the stable communities.

![Figure 5. Two scenarios of data oscillations considered by Heuristic 2](image)

2.1.2. Data Quality Metrics

We employed a set of quality metrics 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 include 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 as follows.
● **Monthly active users (MAU):** the number of devices with at least one sighting for a specific month.

● **Daily active users (DAU):** the number of devices with at least one sighting on a specific day for a specific month.

● **Regularly active users (RAU):** the number of devices with at least seven days of more than ten daily sightings for a specific month.

● **Temporal consistency:** the average number of observed days for RAUs in a specific month.

● **Data frequency:** mean, 25th, 50th and 75th percentile of the average daily number of sightings by RAU devices.

● **Location accuracy:** mean, 25th, 50th and 75th 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.

● **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.

● **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.

● **Device representativeness:** 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.

● **Active local hours:** mean, 25th, 50th and 75th percentile of the average daily number of local hours observed for RAUs.

● **Hourly coverage:** 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.

● **Daily coverage:** 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.

The quality metrics statistics from the one-month raw sighting data panel in 2020 are computed and summarized below. The following numbers are provided to help data users compare the data quality of this raw sighting data panel with that of other similar data sources.

● **Monthly active users (MAU):** 270,601,232 devices, implying a sampling rate of more than 80% on a monthly basis.

---

\(^1\) 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.
- **Daily active users (DAU):** 112,420,233 devices on average during the month, implying an average sampling rate of about 34% on a daily basis.
- **Regularly active users (RAU):** 68,016,290 devices, indicating a sampling rate of more than 20% regarding temporally consistent devices.
- **Temporal consistency:** 24.2 days (the highest possible number of which is 31 days), indicating the level of temporal consistency and coverage of the RAUs.
- **Data frequency:** mean, 25th, 50th and 75th percentiles are 234.4, 72.4, 127.8, and 298.2 sightings per day, respectively, indicating the sighting frequency of RAUs.
- **Location accuracy:** mean, 25th, 50th and 75th percentiles are 49.2, 13.1, 31.1, and 64.6 in feet, respectively, indicating the reliability of location sightings of RAUs.
- **Geographical representativeness (by device):** 0.4, indicating an even geographical distribution of RAUs per population.
- **Geographical representativeness (by sighting):** 0.2, indicating an even geographical distribution of sightings per population.
- **Device representativeness:** 0.6, indicating 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. The team developed a weighting framework to address the uneven distribution.
- **Active local hours:** mean, 25th, 50th and 75th percentiles are 6.4, 2.3, 4.8, and 8.9 hours respectively, indicating high temporal consistency and coverage of the RAUs.
- **Hourly coverage:** 0.2, indicating an even distribution of average daily number of sightings among the 24 hours for RAUs.
- **Daily coverage:** 0.1, indicating 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. So the sighting data generally does not contain any personal information, such as home location, age group, and income level. Such personal information is critical in sample bias correction, weighting, and data expansion. For the national passenger OD data development, the framework only uses sighting data from sample devices whose home locations can be imputed. The sample devices with imputed home locations are 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.

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 evaluates 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 may have occupations like transportation and shipping occupations, whose trips are covered in the national truck OD data products, and cleaning and pest control workers, whose trips are still considered in the national passenger OD data products and whose working profiles are necessary for device-level weighting. 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 weighting and ensure the proper coverage of the national passenger OD data products (see Section 3.4). Those unemployed and those who work from home are categorized as a device without fixed workplace or jobs by the algorithm since there is 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 and developed the framework for imputing home and fixed workplace locations.

Figure 6 introduces the methodology to impute home and fixed workplace locations and worker types.
2.2.1. **Home Location Identification**

To efficiently process the tremendous amount of mobile device location data, the algorithm utilizes 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 identifies the home and workplace at a level-6 geohash zone and then selects 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 identifies 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 favors the home location candidate that is most frequently observed during nighttime and selects it as the home location at level-6 geohash zone level. The first two steps are 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, we 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 are 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 is higher than 0.95.

2.2.2. Fixed Workplace Location Identification

With home location identified, the framework recognizes an individual’s major work location that is not home. Similar to the home location identification, the method considers 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 introduces a temporal similarity ratio between the workplace candidates and identified home location. The motivation is two-fold. First, for the sake of computation efficiency, the home and workplace imputation adopt geohash as the representation of the actual location. If a device dwells 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 may 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 shall not be frequently observed at the same hours. Hence, workplace identification checks the temporal similarity in terms of the specific hours when the home location and workplace candidates are observed to find the most possible workplace location.

For each workplace candidate, the temporal similarity ratio is defined as the ratio between the number of hours when the device was observed both at home and at the workplace candidate and the number of total hours when the device was observed 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 workplace would not be observed at the same hour when the home was observed. 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 is 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.

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 is 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 is 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 identifies the duplicated devices within and between data providers, integrates the sighting data for the duplicated devices, integrates the device and sighting data from all data providers, and creates 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 must meet at least two out of three predefined criteria in terms of device-level data quality metrics. The three metrics are:

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

To ensure the minimum population coverage and avoid privacy concerns for each zone, the sampling rate in all the U.S. counties must be over 5%. Otherwise, all the devices were kept in the counties with sampling rate lower than 5%.

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 is 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 are divided into two groups: devices existing in the previous month’s panel and the remaining devices. The devices existing in the previous panel are favored and thus evaluated with relaxed thresholds. If their device-level quality metrics are higher than the relaxed thresholds, they are kept in the data panel. The remaining devices are 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 time.
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 are ranked by the total number of unique hours observed and the total number of location observations during a month. Devices that share the same home location and the same top five most frequently visited locations (which may include home locations as well) are considered duplicated devices. This is a conservative algorithm, which
ensures that the actual duplicated devices will be captured but may 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 our 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 will be 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 are 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 produces more accurate trip identification results. Figures 8 (a) and (b) show how the tour-based method differentiates 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 is also necessary to identify the true origins and destinations of long-distance trips, especially air trips.

![Figure 8. Tour identification and trip chaining demonstration](image)

The following subsections describe the steps for identifying tours and trips. The algorithm runs on the observations of each device separately.

3.1.1. Home-Based Tour Identification

The algorithm starts with each device’s identified home location (see Section 2.2.1). The home-based tour identification processes 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 will be considered as a home-based tour. Long-distance tours are defined as tours in which a device is observed equal to or more than 50 miles away from its home location. To be consistent with the core travel survey, it is assumed that unless the device is on a long-distance tour, the device starts and ends the trip day at home. In the next step, the sightings of each device are separated into two groups: sightings on short-distance tours and sightings on long-distance tours. Finally, short-distance
tours go through a daily short-distance trip identification and long-distance tours go through a monthly long-distance trip identification.

**3.1.2. Trip Identification for Short-Distance Tours**

It is possible that some sightings do not belong to any trips (i.e., stationary points). For each sighting within the same tour, a recursive algorithm based on the decision tree model is utilized to identify if the sighting is stationary or moving. The decision tree considers 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 has 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 is used to identify if a sighting is recorded on the move, and the distance and time thresholds are used to identify trip ends.

The recursive algorithm checks every sighting to identify if they start a new trip or belong to the same trip as the previous sighting (Figure 9). If the previous sighting is not on a trip (i.e., a stationary sighting), the current sighting starts a trip if it has a speed faster than 3 mph to the next sighting. If the previous sighting is on a trip, the following rules are checked to identify if the current sighting belongs to the same trip, stops the trip, or starts a new trip:

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

The algorithm may identify a local movement as a trip if the device moves within a stay location. To filter out such trips, all trips shorter than 984 ft are 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 follows 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 are filtered. Figure 10 shows how trip identification of long-distance tours works. 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, is applied, but with a larger time threshold of 30 minutes instead of 5 minutes, meaning that a trip ends only if the device stays somewhere for more than 30 minutes. In this step, all the trip ends are identified and named as “secondary stops.” Primary stops are then identified from the secondary stops. Primary stops on a long-distance tour are places where the device stays for a significant amount of time and/or from which the device makes local trips. In order to identify the primary stops, each secondary stop is checked against the following criteria:

- The duration of stay in the secondary stop is longer than two hours and during the stay, the device exits and reenters the secondary stop
- The duration of stay at a location is longer than 24 hours
- The secondary stop is the home location

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

3.1.3.2. Subtour identification

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

3.1.3.3. Trip identification

If a long-distance tour does not have a primary destination or has the same primary destination as the identified workplace, the short-distance trip identification algorithm (with a time threshold of five minutes) is applied to all the sightings in the tour. If a tour has a primary destination different from the fixed workplace, the long-distance trip identification algorithm with a time threshold of 30 minutes is applied to sightings between two different primary stops, and the short-distance trip identification recursive algorithm with a time threshold of 5 minutes is 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 are 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 and in Table 1:

- 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.

Table 1. Literature Review on Travel Mode Imputation Methods

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


According to the literature review done by Huang et al. 2019 and Burkhard et al. 2020, it can be observed that typical features include speed and acceleration (Stenneth et al. 2011; Gong et al. 2012; Brunauer et al. 2013; Nitsche et al. 2014; Xiao et al. 2015; Shafique and Hato 2016; Wang et al. 2017; Dabiri and Heaslip 2018; Broach et al. 2019; Burkhard et al. 2020; Vaughan et al. 2020). Specifically, when the LRI is less than 10 seconds, the speed (speed variation) and acceleration features are more important to differentiate between different travel modes, which can be imputed solely by the data itself. When the LRI is relatively high, for instance, 30 s,
additional features can be added to maintain the same level of accuracy such as real-time transit information (Stenneth et al. 2011), multimodal transportation network (Stenneth et al. 2011; Gong et al. 2012; Burkhard et al. 2020; Breyer et al. 2021), sociodemographic information (Wang et al. 2017; Vaughan et al. 2020), etc.

These methods have been widely applied to impute travel mode with low-LRI sighting data. However, when imputing travel mode from the sighting data, one key issue is that the LRI 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 has 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 is imputed in stages. The air travel mode is firstly imputed based on a heuristic rule calibrated based on ground truth data. Then, an ensemble machine learning model is developed and used to impute land and water 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](image-url)
3.2.1. Air Travel Mode Imputation

As shown in Figure 11, because of its uniqueness in trip features compared to the land and water transportation travel modes, the first step is to impute air trips from the national passenger trip roster. The air trips are 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 is 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 is longer than 50 miles
- The travel time of an air trip is longer than 30 minutes
- The average travel speed of an air trip is faster than 75 mph
- The origin and destination distances to the airport should be 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 is conducted: if the travel time, travel distance, and average travel speed are significantly high and do not belong to any land and water transportation mode, the nearest airports will be assigned for both origin and destination of the trip.

3.2.2. Land and Water Transportation Travel Mode Imputation

After air trips are imputed from the national passenger trip roster, a machine learning model was developed and applied to impute the land and water transportation travel modes for non-air trips, including vehicle (car and bus), rail, and active transportation/ferries (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) are considered for land and water transportation travel mode imputation, as shown in Table 2.

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 are all derived from our sighting data. The multimodal transportation network features are important to distinguish between different land and water transportation travel modes. Here, the distance for each sighting to its nearest rail and bus lines are 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 are also calculated. Also,
the percentage of records within 165 feet (50 meters) of all rail stations or bus stops are calculated for each trip. Those features are used to capture the short stops at rail or bus stations for rail and bus travels since more sightings should 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 can be observed at a further distance from the stations, which relaxes 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) are collected from the Homeland Infrastructure Foundation-Level Data (HIFLD) and U.S. Department of Transportation Bureau of Transportation Statistics.

Table 2. Features for Detecting Land and Water Transportation Travel Mode

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

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 land and water 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 is 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 can achieve 97.1% cross-validation accuracy for land and water 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/ferries (walk, bike, ferry, and other modes).
3.3. Merging Unlinked Trip Segments into Trips

For short- and long-distance tours, we have developed two separate methodologies to link the related unlinked trip segments, 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/ferries 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 active transportation/ferries modes
- Supplementary trip(s) before the access trip
- Supplementary trip(s) after the egress trip

After locating all transit trips, we use different spatial and temporal thresholds when linking 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 should satisfy 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 is 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—are checked to make the access and egress trips complete. The spatial and temporal threshold values applied here are 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 should be smaller than a transfer time threshold value (at most 30 minutes).

The five spatial or temporal threshold values are 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 result 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 are linked, i.e., each subtour will be 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 are linked as one trip with the major land and water transportation mode as the new travel mode and the egress trips leaving from the airport are linked as one trip.

3.4. Worker Type Identification

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

The worker type identification has two major objectives: 1) to identify and remove trips made by professional drivers so that the passenger trip estimates for the population are exclusive of the trips made by professional drivers driving for work which are 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 should be considered in the national passenger OD data products and whose work profiles are necessary for device-level weighting.

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 our algorithm is shown in Figure 13.

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 should be much 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 constitute the basics of our professional driver identification algorithm.

To identify and exclude professional driver trips, the algorithm utilizes 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 uses 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) is added as another threshold. The parameters for the minimum driving hours (three hours) and the minimum number of workdays (nine workdays) are 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 are 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 is frequent and regular travel during the daytime. In the ATUS, more than 25% of the full-time workers are 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 are up to 12%. Therefore, we add one hour of commute time 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. is then adopted as the daytime period for identifying workers without fixed workplaces. The workers without fixed workplaces are 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) are 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 is imputed as work and non-work. The imputation process includes two major parts: data preparation and imputation algorithms. Figure 14 shows the flowchart of the trip purpose imputation method.
3.5.1. Data Preparation

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

- If the trip end and the imputed home location of the corresponding traveler are at the same location, this trip end is labeled as “home”
- If the trip end and the imputed work location of the corresponding traveler (if the traveler has a work location) are at the same location, this trip end is labeled as “work”
- All the other trip ends are labeled as “other”

For the long-distance trips, a two-step model is implemented. First, the long-distance trip ends are matched with the Point of Interest (POI) data. If the traveler stays 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 is labeled as a “Convention Center Staying.” The duration of the stay is 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 is applied. The feature selection for the machine learning model considers features that can be extracted from both the mobile device location data and the travel survey to ensure that the imputation model is 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

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

3.5.2. Imputation Algorithm

3.5.2.1. Short-distance trip purposes

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

3.5.2.2. Long-distance trip purposes

For long-distance trips, the trips are imputed based on the following rules: 1) all long-distance tours labeled with “Convention Center Staying” in the pre-processing step are identified as business tours. 2) the purpose of other long-distance tours is imputed by a machine learning model into one of two categories: business and non-business tours. All the trips in a business tour are considered work trips, and all the trips in a non-business tour are 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, we incorporated a scalable map matching and routing algorithm 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, is first used to find all the roads within 328 ft (or 100 meters) for each sighting. The next step is to construct the complete path between all the sightings snapped to the road networks using routing algorithms. For each sighting, the algorithm first compares 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 is selected as a valid match. Then, the path between the consecutive matched sightings is reconstructed by using the shortest path algorithm based on road length. In the meantime, reasonableness checks are also conducted during the routing process. For each pair of consecutive sightings snapped to the network, the routed distance is first calculated by adding the length of all the road segments routed between the two sightings. Then, two reasonableness checks will be conducted (Newson and Krumm, 2009):

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

If either of these two violations is observed, an incremental approach is 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 do 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 is constructed. Next, the trip distance is 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 are snapped to the rail network and the trip distance is calculated after routing is implemented on the points. Considering that the rail network has significantly fewer links and limited route options compared to the road network, the map matching and routing have higher precision for the rail trips. The trip distance is similarly derived based on the length of the traversed segments for all unlinked rail trips.

### 3.6.2.3. Air travel

For air travel, the geodetic distance of the origin and destination of the trips is used as the trip distance.

### 3.6.2.4. Active transportation/ferries travel

For active transportation/ferries 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, our method relies on the summation of the geodetic distances between all consecutive sightings for each trip.

### 3.6.3. Trip Distance Calculation for Linked Trips

To report the trip distance for linked 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.

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 unweighted national trip roster was generated based on the national device and location data panel. The unweighted national trip roster will then serve as the input to the multi-level weighting and 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 weighting and expanding the unweighted national passenger trip roster and aggregating the weighted national all trip roster into national passenger OD data products.

4.1. Weighting and Data Expansion

Sighting data comes from a non-probability sample of devices and does 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 rate 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.

Weighting and data expansion are two important steps to generate population-representative statistics from a sample. For the NextGen NHTS OD program, a multi-level weighting method was applied (device- and trip-level weighting) to produce OD products that are representative of the entire U.S. population and its corresponding movements (Figure 15).
4.1.1. Device-Level Weighting

For device-level weighting, 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 is device selection. Only devices that have passed the quality check and have home locations identified are considered in the device-level weighting. The monthly sampling rate (the number of devices per population) of such devices at the state level is shown in Figure 16. The overall sampling rate at the national level is 16.1%.
The second step is to address the device sample representativeness bias across different socio-demographic groups through socio-demographics imputation. A decision tree-based machine learning model is 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 categorizes 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 are later used as control totals in the device-level weighting.

Figure 17 shows the framework of the device-level weighting and 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 are further aggregated into 5 groups respectively, resulting in a total of 25 subcategories. The IPF method is then applied at the county level to generate a device-level weight to match the control totals. If a certain county has zero observations in one of the 25 subcategories, the 25 subcategories for that county are aggregated into 9 subcategories to continue the IPF process. If there are still zero observations for one of the nine subcategories, the population-level weights are applied, which are 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.

After the weights 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 weights to represent the 2020 U.S. population.

Figure 17. The framework for county-level iterative proportional fitting
Table 4. Categories Considered in the IPF

<table>
<thead>
<tr>
<th>Initial Categories in the Twenty-Five-Category IPF</th>
<th>Aggregated Categories in the Nine-Category IPF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than 25 years old &amp; less than $25,000/year</td>
<td>Less than 35 years old &amp; less than $50,000/year</td>
</tr>
<tr>
<td>Less than 25 years old &amp; $25,000-$49,999/year</td>
<td></td>
</tr>
<tr>
<td>25-34 years old &amp; less than $25,000/year</td>
<td></td>
</tr>
<tr>
<td>25-34 years old &amp; $25,000-$49,999/year</td>
<td></td>
</tr>
<tr>
<td>Less than 25 years old &amp; $50,000-$74,999/year</td>
<td></td>
</tr>
<tr>
<td>Less than 25 years old &amp; $75,000-$124,999/year</td>
<td></td>
</tr>
<tr>
<td>25-34 years old &amp; $50,000-$74,999/year</td>
<td></td>
</tr>
<tr>
<td>25-34 years old &amp; $75,000-$124,999/year</td>
<td></td>
</tr>
<tr>
<td>Less than 25 years old &amp; $125,000 and more/year</td>
<td></td>
</tr>
<tr>
<td>25-34 years old &amp; $125,000 and more/year</td>
<td></td>
</tr>
<tr>
<td>35-54 years old &amp; less than $25,000/year</td>
<td></td>
</tr>
<tr>
<td>35-54 years old &amp; $25,000-$49,999/year</td>
<td>35-64 years old &amp; less than $50,000/year</td>
</tr>
<tr>
<td>55-64 years old &amp; less than $25,000/year</td>
<td></td>
</tr>
<tr>
<td>55-64 years old &amp; $25,000-$49,999/year</td>
<td></td>
</tr>
<tr>
<td>35-54 years old &amp; $50,000-$74,999/year</td>
<td>35-64 years old &amp; $50,000-$124,999/year</td>
</tr>
<tr>
<td>35-54 years old &amp; $75,000-$124,999/year</td>
<td></td>
</tr>
<tr>
<td>35-64 years old &amp; $50,000-$74,999/year</td>
<td></td>
</tr>
<tr>
<td>35-64 years old &amp; $75,000-$124,999/year</td>
<td></td>
</tr>
<tr>
<td>35-64 years old &amp; $125,000 and more/year</td>
<td>35-64 years old &amp; $125,000 and more/year</td>
</tr>
<tr>
<td>55-64 years old &amp; $125,000 and more/year</td>
<td></td>
</tr>
<tr>
<td>65 years old and above &amp; less than $25,000/year</td>
<td>65 years old and above &amp; less than $50,000/year</td>
</tr>
<tr>
<td>65 years old and above &amp; $25,000-$49,999/year</td>
<td></td>
</tr>
<tr>
<td>65 years old and above &amp; $50,000-$74,999/year</td>
<td>65 years old and above &amp; $50,000-$124,999/year</td>
</tr>
<tr>
<td>65 years old and above &amp; $75,000-$124,999/year</td>
<td></td>
</tr>
<tr>
<td>65 years old and above &amp; $125,000 and more/year</td>
<td>65 years old and above &amp; $125,000 and more/year</td>
</tr>
</tbody>
</table>

4.1.2. Trip-Level Weighting

For the trips identified from mobile device data regarding time of day and trip distance, the major bias in trip estimates is two-fold: 1) the raw sightings of each device may not be complete during the observed time, and 2) the trip identification algorithms may introduce some systematic bias to the imputed trips. The trip-level weighting is 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 is widely discovered between passive data (mainly GPS survey) and survey data estimation. It is 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, the team weighted and calibrated the
national passenger trip roster by mode, departure time, and distance band using the ground truth estimates from the 2017 NHTS. All the following adjustments were applied to the national passenger trip roster that has been adjusted using the device-level weights.

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 state and distance bands were developed based on the average daily mobile device air trip estimates from 10 benchmark days and the average daily reported trips from T-100. The adjustment factors by origin state and distance bands were then applied to the daily mobile device air trip estimates for that month to obtain the final weighted air trip estimates. The weighted monthly totals were summed to the weighted 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. 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. For example, to generate the ground-truth vehicle trip estimates in March 2020, the trend between the monthly VMT estimates in March 2020 and the monthly average VMT estimates during the 2017 NHTS survey period (from April 2016 to April 2017) were applied to the 2017 NHTS vehicle trip estimates. The inflation based on VMT assumes that the trip distance distribution did not change over time. 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 ten benchmark days and the average daily vehicle trip totals from the inflated 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 weighted vehicle trip estimates. The weighted monthly totals were summed to the weighted 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. For each month, we first generated the ground truth rail trip estimates by inflating the 2017 NHTS rail trip estimates with the monthly NTD rail trip trend at the national level. For example, to generate the ground truth rail trip estimates in March 2020, we applied the trend between the NTD rail estimates in March 2020 and the monthly average NTD rail trip estimates during the 2017 NHTS survey period (from April 2016 to April 2017) to the 2017 NHTS rail trip estimates. 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 ten benchmark days and the average daily rail trip totals from the inflated 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 weighted rail trip estimates. The weighted monthly totals were summed to the weighted annual total.

4.1.2.4. Active transportation/ferries travel

Since there are no national-level ground truth data sources that monitor the active transportation/ferries modes travel trends and the COVID-19 pandemic has greatly impacted people’s travel behaviors after February 2020, the team developed the adjustment factors by census division, departure time of day, and distance band based on the average daily mobile device active transportation/ferries mode trip estimates from ten benchmark days in January and February 2020 and the average daily active transportation/ferries trip totals from the 2017 NHTS estimates inflated by the population increase. The adjustment factors by Census division, departure time of day, and distance band were then applied to the daily mobile device active transportation/ferries trip estimates for the entirety of 2020 to obtain the final weighted active transportation/ferries trip estimates. The weighted daily numbers of trips were summed to the weighted annual total.

4.1.3. Trip Distance Distribution Comparison

Figure 18 compares the trip distance distribution for unweighted and weighted passenger trips. The weighting process increased the share of trips shorter than 10 miles, decreased the share of trips between 10 and 100 miles, and barely influence the share of trips longer than 100 miles. The distribution trend among different distance bins was not changed much, which implied that the weighting and data expansion process did not distort the travel trends observed from the sighting data.

![Figure 18. A comparison of distance distribution between unweighted and weighted trips](image-url)
4.2. Aggregating Trip Roster into a National Passenger OD Product

The entire national passenger trip roster and the multi-level weighting were used by the team 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)](image)

After constructing the national passenger OD product, 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. In this process, 82,388 OD pairs were impacted and 922,256 total annual trips were removed.

Also, from the travel demand modeling perspective, there is a general expectation that passenger trips are bi-directional in nature, with similar proportions originating from a zone returning to that zone, particularly across a year’s worth of data. After evaluating the OD pairs for this balance and comparing the imbalance flow at OD pairs, our investigations show a 0.07% imbalance flow at the OD pair level considering the total volume of trips. To address this observation, we used the aggregated information from O to D and from D to O to achieve a 100% balanced OD pairs flow.
5. VALIDATION PLAN

The UMD team developed a rigorous validation plan for 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 document. Product validation targets were based on 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 since the 2020 core NHTS data are not yet available. Both 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) is selected as the metrics 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 the ACS population data with that computed from the 2020 TVT VMT data and the ACS population data since the 2020 HPMS data are not yet available. Both VMT per person estimates had similar spatial trends across different census divisions.

5.1.2. National Air Passenger Trips

The UMD team leveraged the DB1B data and the T-100 data as the calibration and validation data sources for the air trips. 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 air passenger OD data report 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 error is 1.33%, which meets the contract requirement that the error should be within +/- 10%.

5.1.3. National Rail Passenger Trips

The UMD team leveraged the NTD data and the Amtrak ridership 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 were 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 is 8.16%, which meets the contract requirement that the error 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

Besides the aforementioned validation plan, the UMD team examined the national passenger data to ensure that the data have no extreme and unreasonable values in any geography and are logically reasonable.

The team confirmed that the national passenger OD data do not have: (1) “active transportation/ferries” 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 land and water transportation modes, including “vehicle”, “rail”, and “active transportation/ferries” modes; (3) trips between the continental United States and Hawaii by land and water transportation modes, including “vehicle”, “rail”, and “active transportation/ferries” 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


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.


U.S. Department of Transportation, Bureau of Transportation Statistics. 2020 Amtrak Ridership Data.


<table>
<thead>
<tr>
<th><strong>GLOSSARY</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Location Sighting</strong></td>
</tr>
<tr>
<td><strong>Passively Collected Location Data/Mobile Device Location Data</strong></td>
</tr>
<tr>
<td><strong>Geohash</strong></td>
</tr>
<tr>
<td><strong>Monthly Active Users (MAUs)</strong></td>
</tr>
<tr>
<td><strong>Raw Sighting Data Panel</strong></td>
</tr>
<tr>
<td><strong>Daily Active Users (DAUs)</strong></td>
</tr>
<tr>
<td><strong>Regularly Active Users (RAUs)</strong></td>
</tr>
<tr>
<td><strong>Temporal Consistency</strong></td>
</tr>
<tr>
<td><strong>Data Frequency</strong></td>
</tr>
<tr>
<td><strong>Location Accuracy</strong></td>
</tr>
<tr>
<td><strong>Geographical Representativeness (by Device)</strong></td>
</tr>
</tbody>
</table>

---

<sup>2</sup> 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. |
| **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. |
| **Active Local Hours** | Mean, 25th, 50th and 75th percentile of the average daily number of local hours observed for RAUs. |
| **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. |
| **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 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. |
| **Activity Location Identification** | The methodology to identify the most significant locations for each device from a set of activity locations from the sighting data. |
| **Home Location Identification** | The methodology to identify the home location of a device from the sighting data. |
| **Fixed Workplace Location Identification** | The methodology to identify the fixed workplace location of a device if it exists. |
| **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. |
| **Device Deduplication** | The methodology to deduplicate the devices potentially owned by the same individual from the sighting data. |
| **Sighting Data Integration** | The methodology to integrate sighting data from different devices owned by the same individual and from different data providers. |
| 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 weighted 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. |
| Unlinked Trip | The basic unit of analysis for trips. |
| Tour/Home-Based Tour | A sequence of unlinked trips between the departure from and the arrival at one’s identified home location. |
| 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. |
| 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. |
| Secondary Stop | A place where the device stays for more than 30 minutes on a long-distance tour. |
| 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. |
| Primary Destination | The farthest primary stop that is located at least 50 miles away from the home location. |
| Subtour | A segment of a long-distance tour that falls between two primary stops. |
| 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 land and water transportation 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. |
| Trip | Unlinked and linked trips. |
Tour and Trip Identification: The methodology to identify tours and trips from the sighting data.

Trip Chaining: The methodology to merge the unlinked trips into linked trips.

Tour-Based Method: The methodology that first identifies the tours and enables one to consider trip chaining and differentiate between linked and unlinked trips.

Travel Mode Imputation: The methodology to impute the travel mode for unlinked trips from the sighting data.

Trip-Based Approach: The methodology to impute the travel mode for the entire trip.

Segment-Based Approach: The methodology to impute the travel mode for a segment of the trip.

Location Recording Interval: The time duration between the consecutive sightings.

Workers without Fixed Workplace: Workers—such as cleaning and pest control workers—that do not have a fixed workplace.

Professional Driver Identification: Drivers that regularly have driving trips with long trip durations.

Worker Type Identification: The methodology to identify the professional drivers and other workers without fixed workplaces.

Trip Purpose Imputation: The methodology to impute trip purpose for linked trips from the sighting data.

Short-Distance Trip Purpose: The trip purpose for short-distance trips.

Long-Distance Trip Purpose: The trip purpose for long-distance trips.

Business Tour: A tour with business activities as the primary trip purpose.

Non-business Tour: A tour with non-business activities as the primary trip purpose, such as personal recreation.

Trip Distance Calculation: The methodology to estimate the trip distance for each identified trip using the sighting data and road network.

Scalable Map Matching and Routing: The methodology to snap the sighting data to the road network and estimate the path using a routing algorithm.

Device-Level Weighting: The methodology to weight and expand the sample devices to represent the entire population of the U.S.
<table>
<thead>
<tr>
<th>Methodology</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trip-Level Weighting</td>
<td>The methodology to weight and expand the sample trips to represent the travel of the entire U.S. population based on control totals from external ground truth data.</td>
</tr>
<tr>
<td>Multi-Level Weighting</td>
<td>The methodology composed of device- and trip-level weighting.</td>
</tr>
<tr>
<td>Temporal Adjustment Factor</td>
<td>The factor to account for the population growth from 2019 to 2020.</td>
</tr>
<tr>
<td>Socio-Demographic Imputation</td>
<td>The methodology to impute socio-demographic information for each device.</td>
</tr>
</tbody>
</table>