U.S. Department of Transportation
Federal Highway Administration

Next Generation
National Household Travel Survey
National Origin Destination Data

2020 Truck Origin-Destination Data
Methodology Documentation

UNIVERSITY OF MARYLAND

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Next Generation National Household Travel Survey National Origin Destination Data 2020 Truck Origin-Destination Data Methodology Documentation		5. Report Date November 2022 6. Performing Organization Code:
7. Author/s Lei Zhang, Aref Darzi, Yixuan Pan, Mofeng Yang, Qianqian Sun, Aliakbar Kabiri, Guangchen Zhao, Mohammad Ashoori, Saeed Saleh Namadi, Chenfeng Xiong		8. Performing Organization Report No.
9. Performing Organization Name and Address University of Maryland Office of Research Administration 3112 Lee Building College Park, MD 20742-5141		10. Work Unit No. 11. Contract or Grant No. GS-10F-0502N
12. Sponsoring Organization Name and Address Office of Highway Policy Information Federal Highway Administration 200 New Jersey Avenue, SE Washington, DC 20590		13. Type of Report and Period Covered Draft Methods Document. October 2020–November 2022 14. Sponsoring Agency Code

15. Supplementary Notes The Federal Highway Administration’s Task Monitors for this project were Daniel Jenkins, P.E., and Dr. Patrick Zhang, P.E.
16. Abstract This document outlines methods and approaches to develop the origin destination (OD) data covering all aspects of data collection, analysis, and final tabulation for producing the national truck OD data product.

17. Key Words Origin destination data, truck GPS data, truck travel demand, passively collected location data, travel behavior, emerging technologies, data privacy, National Household Travel Survey (NHTS), NextGen NHTS	18. Distribution Statement: Internal draft. NOT for public distribution.

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

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

EXECUTIVE SUMMARY

Federal Highway Administration (FHWA) has launched the Next Generation National Household Travel Survey (NextGen NHTS) program, aimed at establishing a more continuous travel monitoring program with national data products. As part of the NextGen NHTS program, this project produces national multimodal passenger and truck travel Origin-Destination (OD) tables at the national level from passively collected mobile device location data. This document describes the technical approach employed by the University of Maryland (UMD) project team to develop truck OD data for the program. The methodology for producing passenger OD products is documented in a separate project deliverable.

The project was led by UMD. ATRI and INRIX, providers of passively collected truck data, provided data sourcing and analysis support to this project. The UMD team worked with both providers to develop a high-quality truck OD product. The UMD team employed an algorithm to identify all truck trips. A trip linking algorithm was then applied to identify intermediate truck stops and to ensure linking the trip segments without delivery purposes. All sightings of a chained truck trip were then matched to a roadway network for routing and trip distance analyses. A weighting method at the OD and trip level was developed based on truck volume collected at truck counting stations across the nation, with consideration for various types of sampling biases. After proper weighting and data expansion, a national truck trip roster was obtained for the development of national truck OD data products.

In addition, the team implemented a validation plan for the methodology and the final data products to ensure data product quality. National level freight movement datasets, including the Commodity Flow Survey (CFS) and the Freight Analysis Framework (FAF) data, were used for validation and reasonableness checks.

The team is fully committed to enhancing the transparency of all methodological steps in producing truck OD products. This document details the technical approach and validation plan. The source codes of all FHWA-funded computation algorithms for OD data production will also be shared in the public domain, along with national passenger and truck OD data products.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. TECHNICAL APPROACH AND METHODOLOGY OVERVIEW

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

Figure 1 provides an overview of the methodology. ATRI and INRIX provided data sourcing and analysis support to this project. After data preprocessing such as removing sightings with invalid data entries, removing duplicate sightings, etc., the UMD team employed a recursive algorithm to properly identify all trip information, including trip origin, destination, start time, and end time. The team further checked the relationship between identified trips and identified potential trip chains to recover the actual origin and destination of the truck travels. After trip linking, the sightings in a chained trip were snapped to the roadway networks for routing and network trip distance calculation. The result is a national truck trip roster, which becomes the input for the development of truck OD data products at the national level.

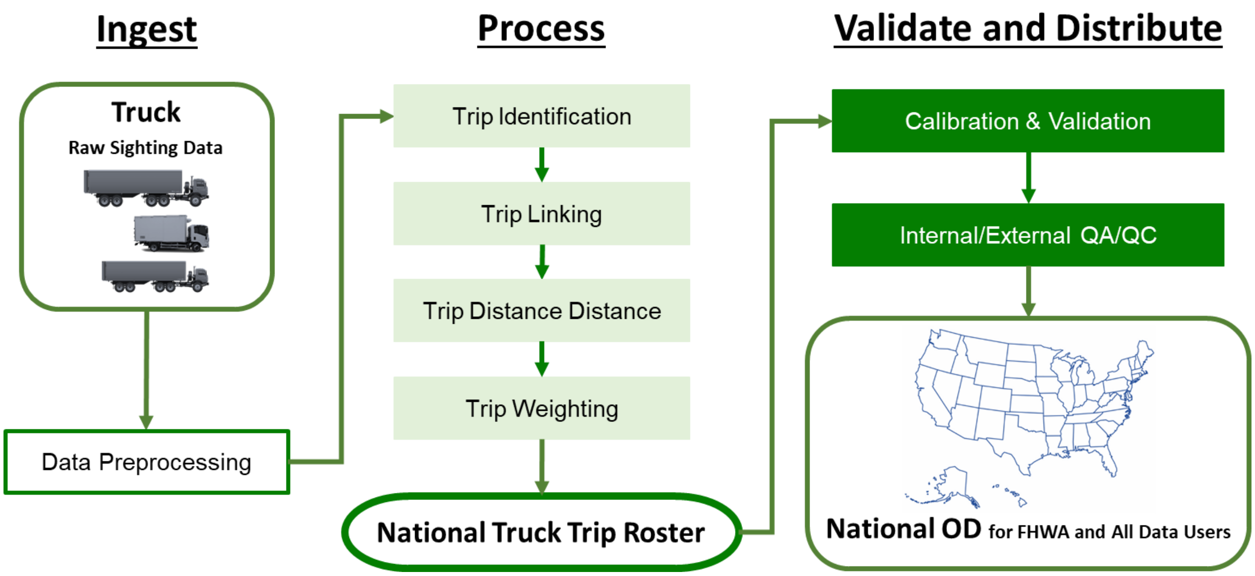


Figure 1. Origin-destination data production flow chart for the Next Generation National Household Travel Survey (NextGen NHTS) truck OD data program

The entire national truck trip roster was used by the UMD team to develop national truck OD products. To expand the observed truck trips from truck roster to the national truck OD product, trips were weighted based on truck traffic count data at the OD level. After truck trip weight calculation, the UMD team calibrated and validated OD products based on the Commodity Flow Survey (CFS) data and Freight Analysis Framework (FAF) data. In addition to that, before delivering national OD data products, a rigorous quality assurance and quality control (QA/QC) procedure was implemented internally by the UMD team.

2. METHODOLOGY

This section describes the methodology for each step in greater detail.

2.1. Data Description and Preprocessing

Passively collected mobile device location data generated from various positioning technologies such as cell phone, Global Positioning System (GPS), and location-based services (LBS), have become increasingly available for transportation planning and operations. For producing truck OD data product, the UMD team received data and analysis support from the two major truck data providers in the United States, ATRI and INRIX. Both datasets represent movements from both urban delivery trucks and long haul trucks. A brief description of each dataset has been provided in the following subsections.

2.1.1. ATRI Sighting Data

ATRI data is collected using in-vehicle GPS devices and details location sightings for a sample[1] of freight trucks from members of the American Trucking Association. The ATRI sighting data provide a consistent anonymized device identifier (ID) for each truck and details the location coordinates in latitudes and longitudes with timestamps. With all the aforementioned merits of GPS data, ATRI data cover stable and complete location sightings for the sample freight trucks from members of the American Trucking Association.

2.1.2. INRIX Sighting Data

INRIX data is generated by the INRIX Traffic Intelligence Network, which globally combines anonymous, real-time GPS probe data from nearly 300 million commercial fleet, delivery, and taxi vehicles, as well as consumer cellular floating vehicle data and GPS-based devices. With exclusive access to one of the nation's largest sources of intercity truck fleet data, INRIX is unique among the industry in its ability to provide high-quality data covering the nation's transportation system. INRIX has experience in providing real-time and historical traffic data and analytics for numerous state DOTs and transportation agencies in the U.S.

The INRIX data includes high-level information, such as provider ID, provider type, vehicle type, and vehicle weight class.

2.1.3. 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, including negative values for latitudes.

[1] The sampling rate for the both truck data providers cannot be provided due to the data agreements in place.

- **Step 2**: remove duplicate sightings considering all data attributes. For completely identical entries, we only keep one entry.
- **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. This step only applies when the coordinates and timestamps are identical in multiple entries.
- **Step 4 (if applicable):** for each device, sort the sightings by timestamps.

2.2. Truck Trip Identification

The NextGen NHTS truck OD passive data product reports on the total annual truck trips. Unlike traditional data sources such as travel surveys, where the data detail the trips made, truck sighting data record the location sightings at specific time intervals while a device moves, stops, stays static, or starts a new trip. Some truck devices are also connected to the vehicle engines, which record the engine status to help identify the truck status. To extract truck trip information from the truck sighting data, trip identification algorithms are used.

The state-of-the-practice methods identify truck trips from passively collected location data using a three-step framework: (1) identify the status of trucks based on time and/or speed thresholds and detect the trips; (2) post-process the detected trips and remove extremely short or unreasonable trips; and (3) validate the trips. Based on a comprehensive literature review and practice scan, the team proposes the following framework to identify truck trips from truck GPS data as discussed in step (1).

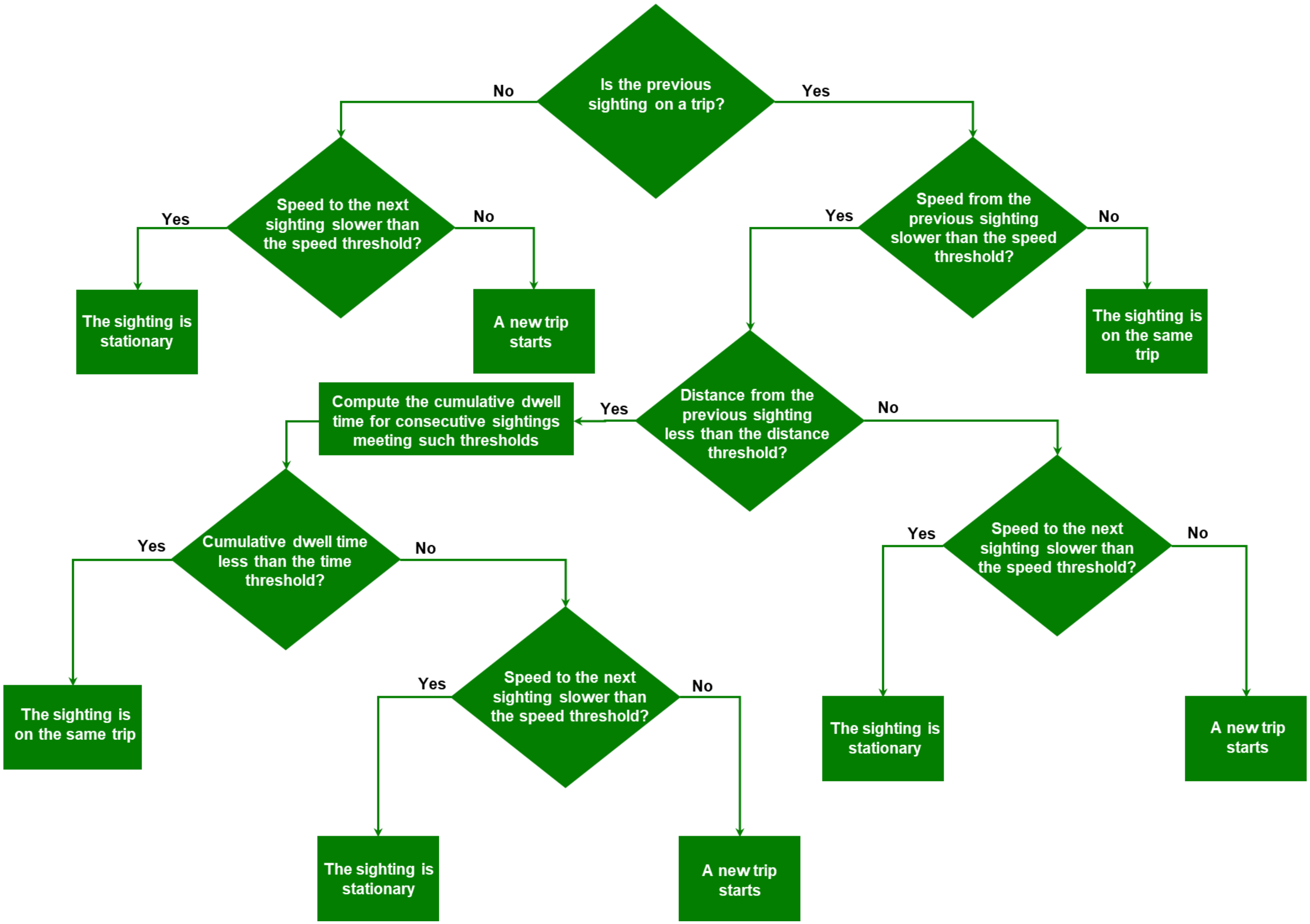


Figure 2. The algorithm for truck trip identification methodology – Step 1 of the framework

To detect trips, the first step is to employ an algorithm to classify the sightings as stationary or moving (Figure 2). To evaluate each sighting's status, the algorithm considers three metrics: (1) the great circle distance, (2) time interval, and (3) speed. For each sighting, the three metrics are calculated for both from the previous sighting to the current sighting and from the current sighting to the next sighting. Therefore, for each sighting, six attributes are considered to distinguish the sighting's status, including speed from, speed to, distance from, distance to, time from, and time to. For these six attributes, three thresholds have been used in the algorithm: a distance threshold of 1800 ft (549 m), a time threshold of 10 minutes, and speed threshold of 2 miles per hour (0.9 meter per second). 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 algorithm checks every sighting to identify if it signals the start of a new trip, if it belongs to the same trip as the previous sighting, or if it terminates the trip (stationary point in Figure 2). 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 2 mph when traveling to the next sighting (speed to). 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 faster than 2 mph from the previous sighting (speed from), the sighting belongs to the same trip as its previous sighting.
- If a sighting has a speed slower than 2 mph from the previous sighting (speed from) and is more than 1800 ft away from the previous sighting (distance from), the sighting does not belong to the same trip as its previous sighting. If the speed to the next sighting is also smaller than 2 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 2 mph from the previous sighting (speed from measure) and is within 1800 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 10 minutes, the current sighting belongs to the same trip; otherwise, it 2) terminates the trip if the speed to the next sighting is less than 2 mph, or 3) starts a new trip if the speed to the next sighting is faster than 2 mph (speed to).

The algorithm may identify a local movement as a trip if the device moves within a small area such as a parking facility location. To filter out such trips, all trips that are shorter than 984 ft (300 m) are removed.

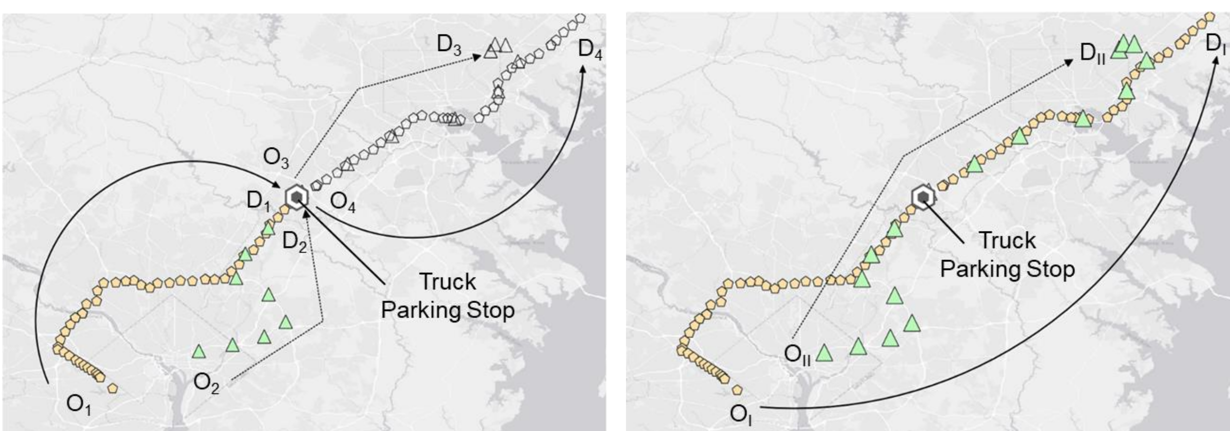
The parameters for distance, time, and speed thresholds are selected based on the literature review and a sensitivity analysis on the semantic information of identified trip ends. The establishments belonging to certain point of interest (POI) types within one mile from the identified trip ends are checked, such as warehouses, grocery stores, gas stations, and truck parking facilities. For each different combination of parameters, the ratio of identified trip ends to such establishments nearby is computed and compared.

The identified truck trips may belong to a long-distance truck trip with necessary intermediate stops for purposes like sleeping, food, and gas. Section 2.3. further evaluates the relationship between detected truck trips and constructed truck trip chains to discover the primary OD pairs for long-distance truck trips.

2.3. Truck Trip Linking

Trip linking is a common process to link basic trip segments into trip chains for different applications in both passenger and freight travel demand analysis, e.g., forming home-work-based tours and long-distance truck tours (Zhang et al., 2012; USDOT, 2017; Xiong and Zhang, 2013; Lu et al., 2012; Zhang et al., 2020). Following the truck trip identification, a truck trip linking framework is developed based on the following motivations: (1) the truck GPS data sources include input data without consistent device IDs, as described in Section 2.1, and (2) actual delivery stops and intermediate stops for non-delivery purposes, such as refueling and sleeping, should be distinguished. The national truck OD products are estimates of truck trips with delivery purposes.

An illustration of truck trip linking is shown in Figure 3. Figure 3(a) shows sightings from four identified trips with four distinct dynamic device IDs. For each trip, one of the trip ends is observed at a truck parking stop. The truck trip linking algorithm evaluates the similarity of the trip characteristics (such as average travel speed) and the temporal intervals between trips ending at and departing from the truck parking stop (the temporal threshold is described in Section 2.3.2). As a result, two linked trips are identified from the four individual trips and therefore are identified as made by two unique trucks: 1) O_I–D_I linked from O_1–D_1 and O_4–D_4 and 2) O_{II}–D_{II} linked from O_2–D_2 and O_3–D_3, as shown in Figure 3(b).



(a) Four identified trip trajectories from four dynamic device IDs

(b) Two linked trip trajectories considered from two unique trucks

Figure 3. Trip trajectories before and after truck trip linking

The overall methodology for truck trip linking is shown in Figure 4 based on the points of interest (POI) data and trip characteristics. The truck trip linking algorithm first considers the identified trips starting from or ending at relevant establishments, including truck parking facilities, gas stations, auto service shops, major arterials, etc. Second, for matching and linking candidate trips related to the same establishment, we jointly evaluate the truck characteristics (such as the data provider and weight class); the temporal differences between trips ending and starting here and their trip characteristics, such as tracking frequency (the number of sightings per minute); OD direction; and average speed. Next, the initial trip pairs matched by establishments with selected POI types are further linked to recover the complete multi-day truck trip itinerary. Finally, to avoid over-linking, a post-processing step is performed to break down round trips.

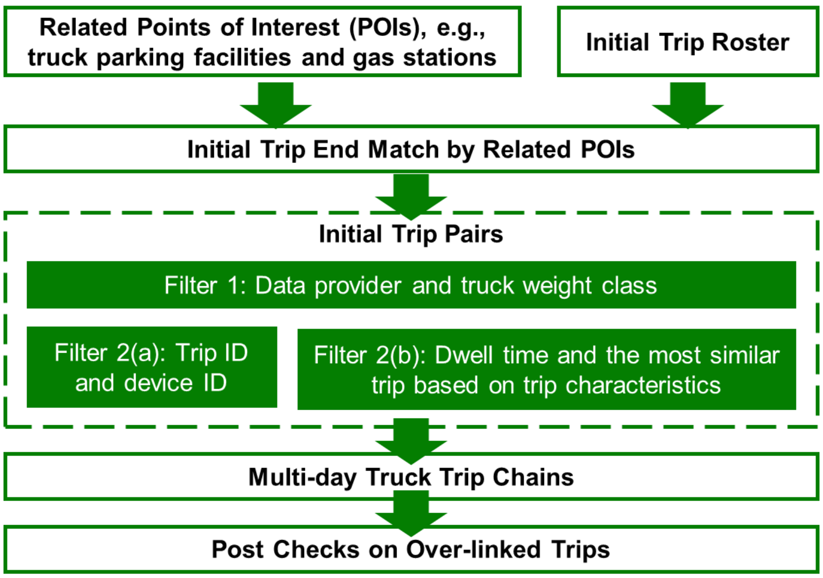


Figure 4. Methodology for truck trip chains formation

2.3.1. Initial Trip End Match by Related POIs

The vicinity of all the identified trip ends is searched for establishments related to trucks' non-delivery stops, including truck parking facilities (USDOT, 2014), gas stations, auto service shops, and major arterials (illustrated in Figure 5). The search radius is 2640 ft (0.5 mile) for truck parking facilities, 1320 ft (0.25 mile) for gas stations and auto service shops, and 200 ft for major arterials. For each category of establishments, the closest one within the search radius is matched with the trip end. Therefore, each trip has at most one matched establishment in each category.

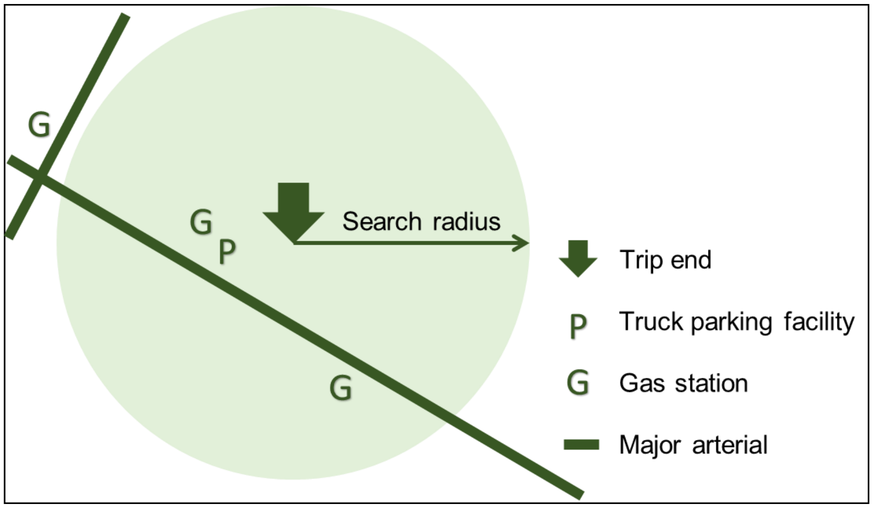


Figure 5. Illustration of trip end match process

2.3.2. Initial Trip Pair Match by Related POIs

For trips with consistent device IDs, the dwell time in the related establishments is evaluated to distinguish between delivery and non-delivery stops. If the truck stays near a truck parking facility for less than 24 hours, near a gas station or auto service shop for less than 2 hours, or near a major arterial for less than 2 hours, the stop is considered a non-delivery stop. The temporal thresholds are selected based on the dwell time summary of the sample trucks to include most such trips as candidates for initial trip pair match. They are also long enough to include most trips with non-delivery purposes, such as sleeping and food. The previous trip arriving at the stop and the next trip departing from the same stop are linked as an initial trip pair.

For trips without consistent device IDs, the trip pair is formed by each related establishment. For each trip arriving at a related establishment (referred to as the target trip), the trips departing from the establishment within a certain timeframe are considered possible next trips by the same truck (referred to as the candidate trips). The timeframe for different types of POIs is defined as follows:

- less than 24 hours after the arrival time of the target trip for truck parking facilities;
- less than 2 hours after the arrival time of the target trip for gas stations and auto service shops; and
- less than 2 hours after the arrival time of the target trip for major arterials.

Among the candidate trips, the trip with the most similar trip characteristics in terms of sighting frequency, travel direction, average speed, and maximum speed to the target trip is selected to form an initial trip pair with the target trip.

2.3.3. Multi-Day Trip Chain Formation from Initial Trip Pairs

For each trip that is not the second trip in an initial trip pair, a seven-day window from the departure day is searched for consecutive initial trip pairs. The seven-day window for linking purposes has been selected to discover very long-distance trip cases and ensure the completeness of such cases in our product. Additional algorithm has also been developed to prevent the over-linking of the trips within the seven-day window.

The consecutive initial trip pairs are chained together to form the final multi-day trip chain. For example, in the seven-day window, the first trip, i.e., Trip A, forms an initial trip pair with Trip B, Trip B forms an initial trip pair with Trip C, and Trip C does not form an initial trip pair with any later trips. Trip A, B, and C are chained as a final trip chain. The origin of Trip A and the destination of Trip C are considered the origin and destination of the chained trip.

2.3.4. Post Checks on Over-Linked Trips

In the previous step, the algorithm is designed to link all the potential trips with non-delivery purposes. The aggressive algorithm may also mistakenly link some trips with delivery purposes. Therefore, a final post-processing step is developed to address the potential over-linking issue. Two types of linked trips are considered as potential over-linked trips: 1) the linked trips with many intermediate stops, i.e., more than five, and 2) the linked trips with a high detour factor larger than two, which is the ratio between the travel distance and the airline distance between the trip origin and destination (OD distance). The detour factor threshold is selected based on the regular roadway network geometry.

The final step post-checks the OD distance change as more original trips are linked together (illustrated in Figure 6). For a regular linked trip, the OD distance should increase or only have a minor decrease once (percentage decrease less than 20%) when a new trip segment is linked. Figure 6(a) shows a regular linked trip consisting of four original trips with increasing OD distance, i.e., O_1–D_1, O_1–D_2, O_1–D_3, and O_1–D_4, as more trips are linked. Figure 6(b) shows an over-linked trip consisting of six original trips with increasing OD distance for the first three trips, i.e., O_1–D_1, O_1–D_2, and O_1–D_3. However, there is a significant decrease from O_1–D_3 to O_1–D_4, which indicates a delivery trip end in D_3. The over-linked trip is then separated as two linked trips, i.e., O_1–D_3 and O_4–D_6, where the second linked trip O_4–D_6 has increasing OD distance as more trips are linked.

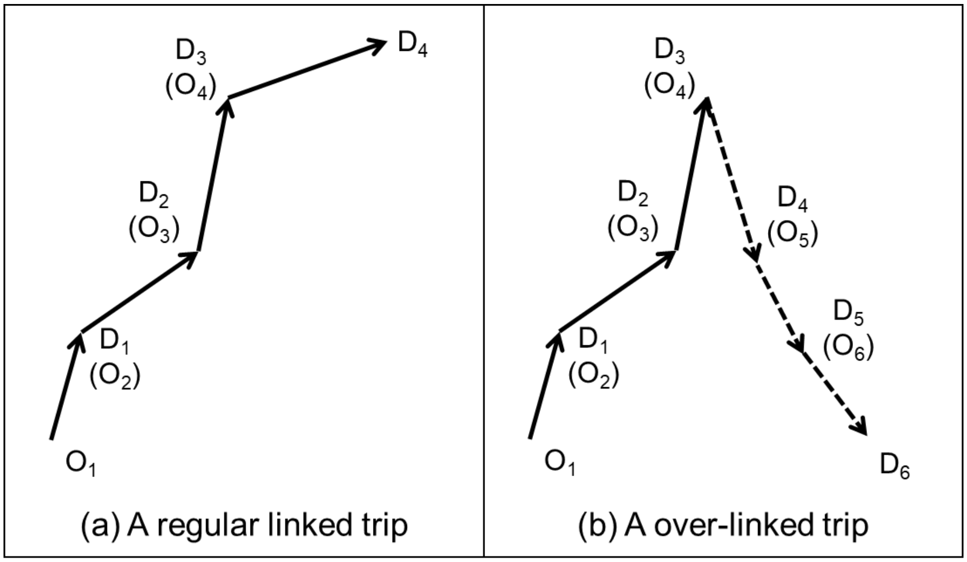


Figure 6. Illustration of post-checking on over-linked trips

Before post-processing, the percentage of potential over-linked trips with a detour factor greater than 2 is 51%. After post-processing, 95% of the revised linked trips from those over-linked trips have a detour factor smaller than 2.

2.4. Trip Distance Calculation

To produce reliable truck trip statistics by distance bands for each OD pair and truck vehicle miles traveled (VMT) statistics for validation purposes, 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 which again has been shown would result in underestimation of trip distance as drivers often fail to follow the shortest paths and the actual paths taken can be longer than the shortest paths. In this project, we incorporated a scalable map matching and routing algorithm to reconstruct the path of all truck trips and then calculate their trip distances based on the observed travel routes. The detail of map matching and routing algorithms used for trip distance calculation for single and chained trips is described below.

2.4.1. Map Matching and Routing for Trip Distance Calculation

The UMD team has developed and implemented a computationally efficient method for snapping sightings to routable transportation networks to reconstruct the complete travel routes for each trip. We implement the spatial index method and shortest path routing algorithm to first find all the roads within 328 ft (or 100 meters) for each sighting, compare the travel direction between the sighting and the road travel direction to snap the sighting to the matched road, and then construct the complete path between all the sightings snapped to the road networks. At the same time, we perform two reasonableness checks on our method:

- If the routed distance is greater than the cumulative distance, calculated by summing the airline distances between all consecutive sightings, between the two observed snapped to the network by 1.24 miles or more, we consider the route as invalid and in need of revision.
- The travel time on these links is calculated based on the timestamp difference between 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), we consider one of the two sightings is matched to the wrong link.

If either of these two violations is observed, we conduct an incremental approach by randomly removing one of the sightings, conduct the routing with the previous/next sighting snapped to the network, and examine the distance and travel speed until they do not violate the 1.24-mile threshold or the 112-mph threshold (Newson and Krumm, 2009).

After implementing the map matching and routing algorithm for all truck 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.

2.4.2. Trip Distance Calculation for Chained Trips

To report trip distance for chained trips, the calculated distances for all the trips that are chained together are summed with gap distances between the trips in each initial trip pair added. The gap distance is the distance from the end of the first trip and the start of the second trip in each initial trip pair.

2.5. Weighting and Data Expansion

Although the truck GPS data have a large temporal and spatial coverage, they are originally from a sample of all U.S. trucks. Additionally, the trips identified from the sightings are a sample of the trips made by the sample trucks due to occasional GPS signal interruption. Therefore, data expansion is necessary to expand the sample trips to population-level truck travel demand. In the meantime, there may exist potential spatial and temporal biases related to different levels of roadways, urban and rural roadways, and monthly sighting data volumes. Such biases must be addressed by a weighting treatment. Both weighting and data expansion need information from external ground truth data, i.e., monthly truck sensor counts from FHWA's Travel Monitoring Analysis System (TMAS).

After truck trip identification, trip linking, and distance calculation, the sample truck trip roster is ready for weighting and expansion. The designed weighting and expansion framework addresses the aforementioned biases and expands the sample trips to population-level truck travel demand simultaneously. The process is independently conducted on a monthly basis and assigns each sample truck trip with a weight.

Our sample trips are divided into three groups and different weighting strategies are applied, respectively (Figure 7):

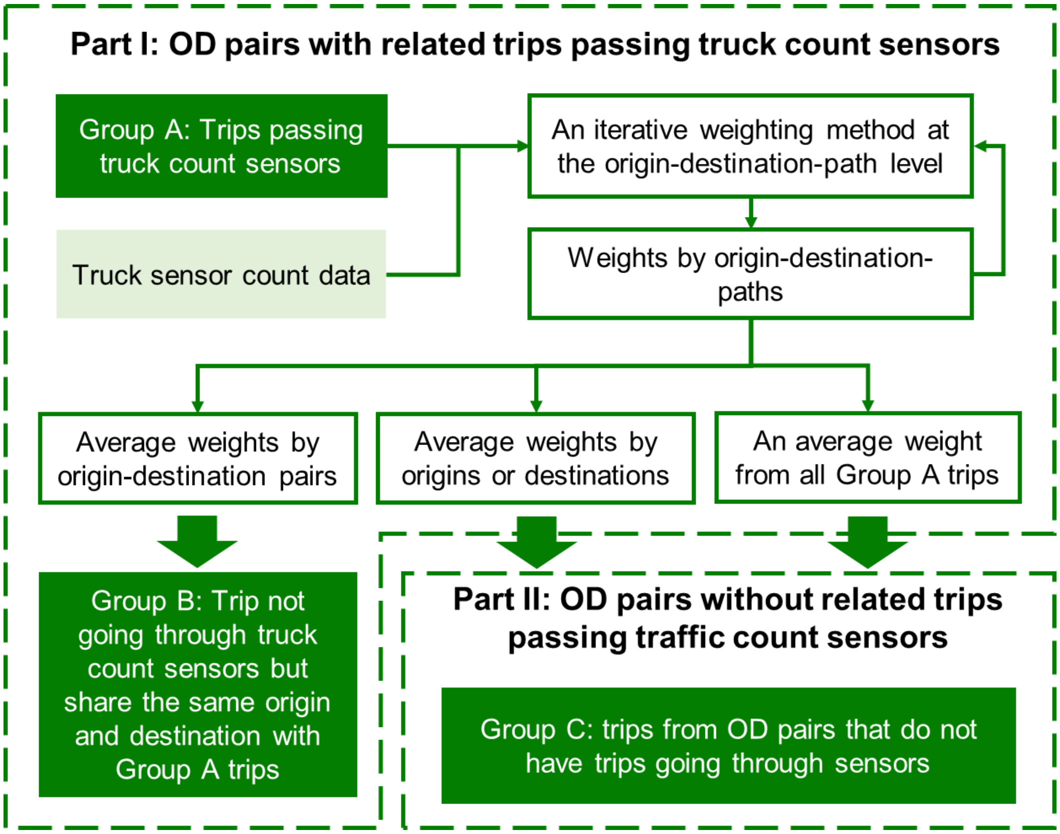


Figure 7. Methodology for truck trip weighting and expansion

- Group A consists of trips going through truck count sensors; an iterative weighting method is conducted to expand these sample trips to match the truck sensor counts.
- Group B consists of trips not going through truck count sensors but that share the same origin and destination with other trips going through sensors, i.e., trips in Group A; the average weight derived from trips going through truck count sensors with the same origin and destination is applied.
- Group C consists of trips from OD pairs that do not have any trips going through truck count sensors; the average weight derived from Group A trips is utilized (more details are described in the following paragraphs).

The iterative weighting method for trips from Group A works as follows.

1. An initial expansion factor, i.e., the ratio of total sensor counts to total sample traffic going through sensors, is applied as the initial trip weights for computation efficiency.
2. An iterative weighting method is implemented at the origin-destination-path (OD-path) level. An OD-path is defined as a sequence of sensors that a truck trip passes

chronologically while traveling between a certain OD pair. During the iterative weighting process, the trips belonging to the same OD-path are assigned the same weights.

3. For each iteration and each sensor, the difference between the adjusted sample traffic and the observed sensor counts for each sensor is reduced (with a step size of 10%) by proportionally adjusting the weights of all the OD-paths going through this sensor. The proportion is the initial volume distribution among those OD-paths at a sensor.
4. Each OD-path may receive multiple adjustments from different sensors. The mean of such adjustments is assigned to each OD-path (with a marginal control of +/- 10%). Such an adjustment process is iteratively conducted until the weighted mean squared error across all sensors is less than 15%.

The iterative weighting method for trips from Group B works as follows. Based on the OD-path weights derived from Group A trips, the average weight for each involved OD pair is computed and later applied to Group B trips with the same origin and destination.

Some origin-destination pairs do not have any sample trips going through sensors. Such sample trips, i.e., Group C trips, are weighted by the following strategies.

1. First, for each origin (or destination), the average weight is computed as the weighted arithmetic mean from the weights of all the Group A's OD-paths departing from the same origin (or arriving at the same destination) and the initial volume distribution of those origin-destination-paths. Then,
2. if the average weight exists for both the trip origin and destination, the squared root of the two average weights is adopted as the trip weight;
3. if only one average weight exists for either the trip origin or destination, the only average weight is adopted as the trip weight; and
4. if no average weight exists for either the trip origin or destination, the average weight from all Group A trips is adopted as the trip weight.

The descriptive statistics on utilized truck count sensors and OD pairs in different scenarios are summarized in Table **1**. Most OD pairs (more than 97% in most months, as shown in Table **1**) have trips going through sensors that are directly weighted using an iterative weighting method based on sensor count data. Among the other OD pairs, most of them have average weight from both origin and destination (more than 1%); very few OD pairs (less than 0.005%) have at least one average weight from origin or destination; very few OD pairs (less than 0.002%) lack an average weight from either origin or destination, the corresponding trips of which are assigned the average weight from Group A trips.

Table 1**. Summary of monthly matched sensors and OD pairs in different cases**

Month	OD pairs with trips passing sensors	OD pairs without trips passing sensors			
		Average weight exists for both origin and destination	Average weight only from origin	Average weight only from destination	Average weight from trips passing sensors
1	98.0%	2.0%	0.002%	0.000%	0.001%
2	97.6%	2.4%	0.000%	0.000%	0.001%
3	98.0%	2.0%	0.002%	0.000%	0.001%
4	97.8%	2.2%	0.000%	0.000%	0.001%
5	97.7%	2.2%	0.001%	0.000%	0.001%
6	97.9%	2.1%	0.000%	0.000%	0.001%
7	98.4%	1.6%	0.005%	0.000%	0.001%
8	98.6%	1.4%	0.000%	0.000%	0.001%
9	98.6%	1.4%	0.000%	0.000%	0.001%
10	98.3%	1.7%	0.000%	0.000%	0.001%
11	98.2%	1.8%	0.002%	0.000%	0.002%
12	93.4%	6.6%	0.000%	0.000%	0.002%

The weighted truck traffic and the truck sensor counts by sensor location are shown in Figure 8. In most sensors, the weighted truck traffic matches the observed truck sensor counts well. It is observed that the error by sensor negatively correlates with the observed truck count; i.e., sensors with higher observed truck volumes have smaller errors. A similar observation can be found in Figure 9. Figure 9 summarizes the average error from sensors by different monthly volume group and the percentage share of the number of sensors belonging to different monthly volume groups. The highest average error comes from the group of sensors with a monthly total observed truck count of no more than 0.02 million (48.73% of total sensor records). In comparison, all other sensors with a monthly total count of more than 0.02 million reach an average error of 6.90%. The potential cause of the observation can be related to the lower chance of observing adequate sample size in cases with fewer traffic volumes and also the termination of the iterative algorithm, which is determined by the weighted mean squared error across all sensors in each OD-path. The logic forces the sensors with higher volumes to have small enough errors but does not have complete control over other sensors with lower volumes.

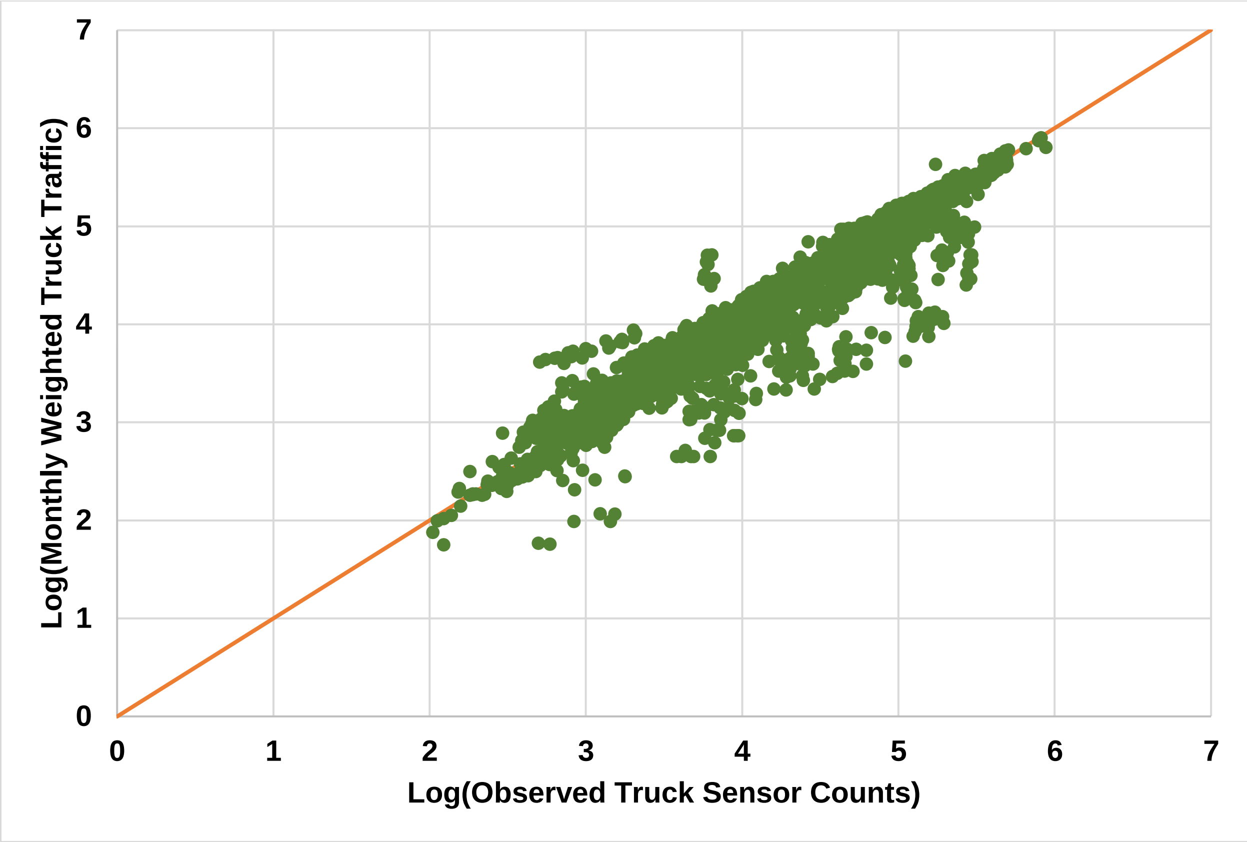


Figure 8. Comparing the monthly weighted truck traffic and observed truck sensor counts

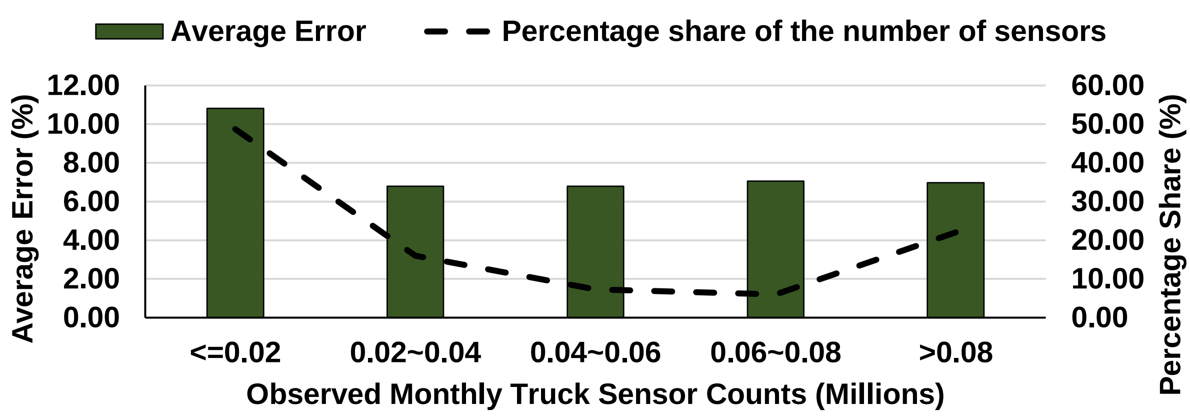


Figure 9. Average error of sensors from different groups of monthly-observed truck counts

Figure 10 and Figure 11 show how the weighting and expansion process mitigates the spatial and temporal biases by state and month. For unweighted trips, the trip ratio between total sensor counts and the total number of unweighted truck trips passing those sensors is computed for each state (minimum: 1.99; maximum: 18.85; SD: 3.20). For weighted trips, the trip ratio between total sensor counts and total number of weighted truck trips passing those sensors is computed for each state (minimum: 0.88; maximum: 1.33; SD: 0.09). Then the ratio between the trip ratio for each state and the average trip ratio of all states is calculated for both unweighted trips and weighted trips and visualized in Figure 10. It shows that the uneven distribution of differences between unweighted trips and truck sensor counts by state has been significantly mitigated.

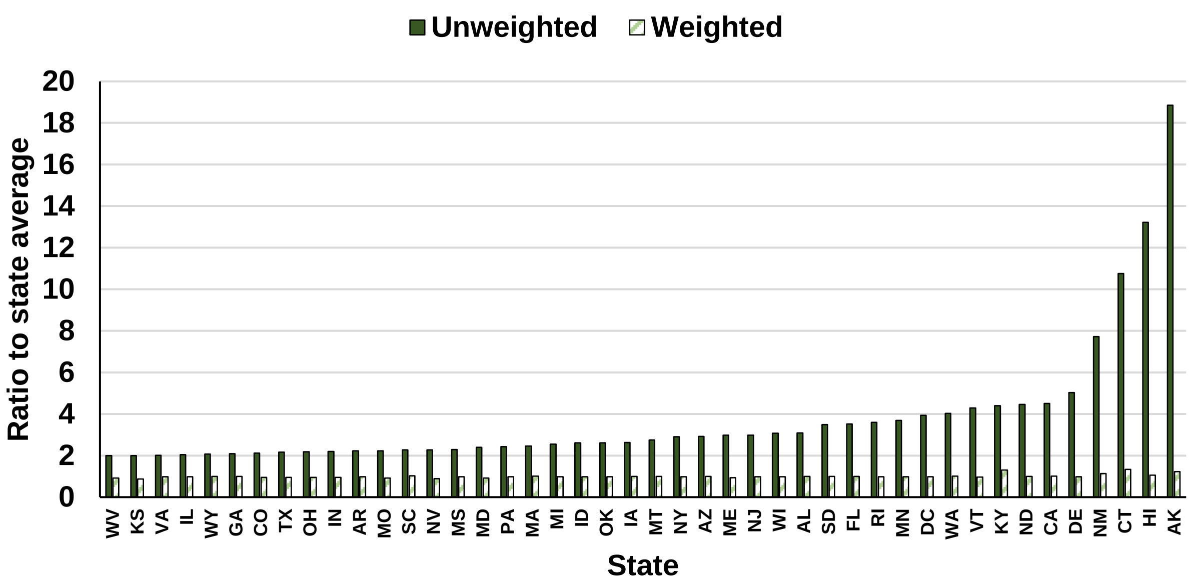


Figure 10. A distribution comparison of differences between truck sensor counts and GPS data trip estimates by state before and after weighting

Similarly, for unweighted trips, the trip ratio between total sensor counts and total number of unweighted truck trips passing those sensors is computed for each month (minimum: 2.47; maximum: 3.62; SD: 0.31). For weighted trips, the trip ratio between total sensor counts and total number of weighted truck trips passing those sensors is computed for each month (minimum: 0.98; maximum: 1.00; SD: 0.01). Then the ratio between the trip ratio for each month and the average trip ratio of all months is calculated for both unweighted trips and weighted trips and visualized in Figure 11. It can be observed that the uneven distribution of differences between unweighted trips and truck sensor counts by month has been significantly mitigated.

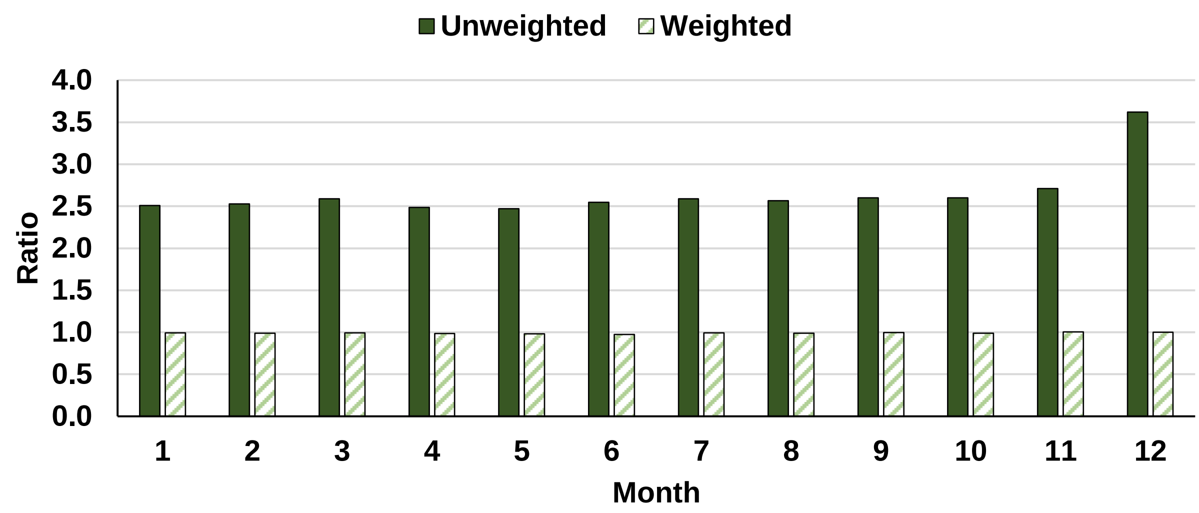


Figure 11. Comparing the ratio of observed sensor counts to truck rosters before and after weighting across the months

In addition to these two biases, other potential biases are related to travel distance, travel time, truck weight class, etc. Due to the unavailability of ground truth data, such biases are not evaluated and shall be explored in the future.

Figure 12 compares the trip distance distribution for unweighted and weighted truck trips. The weighting process has significantly increased the share of trips shorter than 25 miles and decreased the share of trips longer than 50 miles. The distribution trend among different distance bins is not changed much.

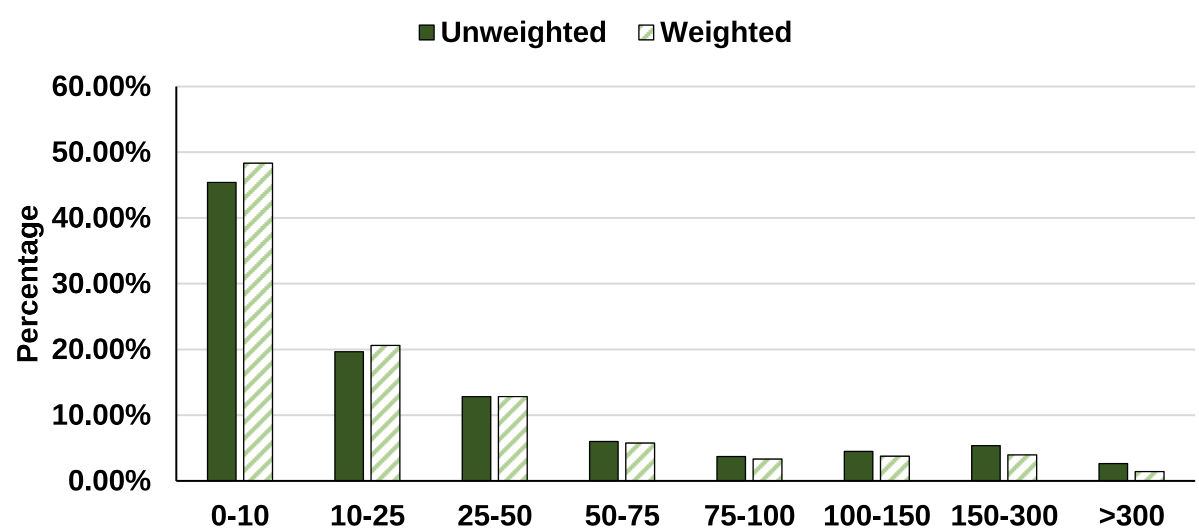


Figure 12. A comparison of distance distribution between unweighted and weighted truck trips

2.6. Aggregating Trip Roster Results into National OD Products

The weighted national truck trip roster is used to develop national truck OD products. In addition to the annual trip totals, the absolute trip numbers and percentage shares by trip distance are also generated.

After constructing the national truck OD product by aggregating the weighted national truck trip roster, several additional checks were employed. The balance of the total annual trip flow between different directions of the same OD pairs is checked. The comparison of trip flow is summarized in Table 2.

Table 2. OD Balance Comparison

Range of return trips to outbound trips	# OD pairs	% Pairs	# Outbound Trips	# Return Trips
<10% returns	20,321	17.75%	2,846,472	85,585
10-<25% returns	4,322	3.78%	6,515,355	1,247,515
25-<50% returns	18,085	15.80%	64,013,919	26,720,319
50-<75% returns	31,882	27.85%	451,058,095	298,825,055
75-100% returns	39,854	34.82%	1,782,438,313	1,602,327,737
Total	114,464	100%	2,306,872,154	1,929,206,211

Based on the summary provided in Table 2, the majority of the truck trip volume (more than 98.5%) occurred in OD pairs with a high range of return trips to outbound trips. Because of the unique pattern of truck trips, the UMD team did not take any further actions to balance the flow of OD trips.

3. VALIDATION PLAN

The UMD team developed a rigorous validation plan for the algorithms and the final data products to ensure product quality and transparency. The validation focuses mainly at aggregate levels because there are no ground truth data for national truck trips. The truck OD products are compared with available freight movement datasets, such as the Commodity Flow Survey (CFS) data (USDOT BTS, 2015) and the Freight Analysis Framework (FAF) data (USDOT, 2019), for reasonableness check.

The team conducted both an internal and an external independent quality assurance and quality control (QA/QC) of the national truck OD data products. Both the internal and external QA/QC follow a similar procedure assessing the key elements of the products, as outlined in this section.

3.1. Comparison with Available Freight Movement Data

As a major data source for freight movement, the Freight Analysis Framework (FAF) is produced based on the Commodity Flow Survey (CFS) and other related data sources through a partnership between the Bureau of Transportation Statistics (BTS) and Federal Highway Administration (FHWA). FAF truck tonnage estimates and corresponding truck traffic estimates were mainly utilized for comparison purposes since the truck samples surveyed by CFS and included in GPS data have distinctive features. In addition, the FAF regional zone system is different from the NextGen NHTS MSA zone system, which only enables a comparison based on limited number of zones. There, the FAF zones that can be nested with the NextGen NHTS MSA zones are selected, for which the spatial correspondence between the two zone systems was developed and used for the following zone-related QA/QC checks. The QA/QC checks for the national truck OD data included but were not limited to: (1) cargo tonnage per truck trip by OD pair; (2) top and bottom FAF zones regarding the percentage of intrazonal truck trips and the percentage of intrazonal truck trip estimates from the NextGen NHTS OD products; (3) top and bottom FAF zones regarding the percentage of interzonal truck trips and the percentage of interzonal truck trip estimates from the NextGen NHTS OD products; (4) Pearson correlation between the FAF tonnage and the NextGen trip estimates by OD pair; (5) Pearson correlation between the FAF converted traffic and the NextGen trip estimates by OD pair.

To check the reasonableness of the NextGen NHTS truck OD products, Freight Analysis Framework (FAF) Version 5.4. estimates for 2020 are employed as the validation data source. In addition to tonnage volume directly from the data, the converted truck traffic from FAF 4.5.1. 2017 data following the methodology provided by FAF are used. The Pearson correlation coefficient between FAF tonnage (and converted traffic) and NextGen NHTS truck trips by OD pair is 0.9, indicating a high correlation and similar trend from the two data. Similarly, the Pearson correlation between the FAF converted traffic and the NextGen truck trip estimates by OD pair is 0.9.

From more than 100 FAF zones, 63 reconstructed FAF zones can be nested with 321 zones from the NextGen MSA zone system. In total, there are 3,035 OD pairs with reported tonnage volume from reconstructed FAF zone system and 209 OD pairs with annual converted traffic of more than 100,000 trips. The average tonnage per trip is calculated for those 209 OD pairs using the FAF tonnage and converted traffic and the NextGen OD trip estimates. A comparison of the descriptive statistics is shown in Table 3. In general, the tonnage per trip estimates from the NextGen OD products are significantly smaller than those from FAF. NextGen OD products include more urban delivery trucks and light-duty trucks, which leads to the discrepancy. Meanwhile, about 10% OD pairs have larger tonnage per trip estimates from the NextGen OD products, many of which are long-distance OD pairs. The potential cause is that the long-distance truck trips have intermediate stops identified from the GPS data and some of them may not be linked by the trip linking algorithm because of reasons such as extra-long dwell time and so forth.

Table 3. A Comparison of Descriptive Statistics of Tonnage per Trip between NextGen Truck OD Products and FAF

Variable	NextGen NHTS Truck Data	FAF Data
Count	209	209
Mean	11.1	17.4
Standard Deviation	26.4	1.8
0%	0.4	11.7
5%	0.7	14.1
10%	0.9	15.0
25%	1.6	16.2
50%	3.2	17.5
75%	6.5	18.6
90%	23.0	19.6
95%	52.3	19.9
99%	128.1	21.1

The team calculates the ratio of interzonal trips and all trips for each zone, and compares that ratio with Freight Analysis Framework (FAF) data for each zone. Table 4 reports the distribution of the two ratios. Overall the distributions of the two ratios are similar to each other.

Table 4. The Percentage of Interzonal Trips in All Trips, a Comparison between NextGen NHTS Truck Data and Freight Analysis Framework (FAF) Data

Variable	NextGen NHTS Truck Data	FAF Data
Count	57	57
Mean	10.2	19.4
Standard Deviation	10.1	16.5
0%	0.0	0.0
5%	0.0	0.0
10%	0.0	0.0
25%	2.4	5.0

50%	7.3	16.8
75%	16.3	29.1
90%	22.8	40.4
95%	31.7	50.4
99%	35.9	57.9

3.2. Reasonableness Check

Besides the aforementioned comparison, the national truck OD data was examined to ensure that the data had no extreme or unreasonable values in any geography and were logically reasonable. Such checks for the national truck OD data included: (1) interzonal trips for those zones that are inaccessible to other zones through trucks (such as Alaska and Hawaii); (3) proportion of intrazonal to interzonal truck trips for each OD pair; (4) truck trip distribution by origin and destination; (5) interzonal truck trip distribution by OD pair; (6) ratio of truck to passenger trips for each OD pair; (7) national-level ratio of truck to passenger vehicle miles traveled; (8) pickup or delivery locations' reasonableness check for some selected zones.

4. REFERENCES

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

Lu, Y., Zhu, S., & Zhang, L. (2012). A machine learning approach to trip purpose imputation in GPS-based travel surveys. 4th Conference on Innovations in Travel Modeling, Tampa, Florida.

U.S. Department of Transportation, Freight Management and Operations. (2015). Jason's Law 2015 Truck Parking Survey and Assessment. https://ops.fhwa.dot.gov/freight/infrastructure/truck_parking/index.htm.

U.S. Department of Transportation, Federal Highway Administration. (2017). 2017 National Household Travel Survey. 2017 NHTS Trip Chaining Dataset. https://nhts.ornl.gov/assets/2016/download/TripChain/Documentation.pdf.

U.S. Department of Transportation, Federal Highway Administration. (2021). Travel Monitoring Analysis System (TMAS).

U.S. Department of Transportation, Bureau of Transportation Statistics and Federal Highway Administration. (2019). Freight Analysis Framework. https://ops.fhwa.dot.gov/freight/freight_analysis/faf.

U.S. Department of Transportation, Bureau of Transportation Statistics. (2015). 2012 Commodity Flow Survey United States. https://www.bts.gov/archive/publications/commodity_flow_survey/2012/united_states/index.

Xiong, C., & Zhang, L. (2013). Deciding whether and how to improve statewide travel demand models based on transportation planning application needs. Transportation Planning and Technology, 36(3), 244-266.

Zhang, L., Southworth, F., Xiong, C. and Sonnenberg, A., (2012). Methodological options and data sources for the development of long-distance passenger travel demand models: a comprehensive review. Transport Reviews, 32(4), pp.399-433.

Zhang, L., Ghader, S., Darzi, A., Pan, Y., Yang, M., Sun, Q., Kabiri, A. and Zhao, G. (2020). Data Analytics and Modeling Methods for Tracking and Predicting Origin-Destination Travel Trends Based on Mobile Device Data. Federal Highway Administration Exploratory Advanced Research Program.

GLOSSARY

Sighting	A sighting is generated from a truck's in-vehicle Global Positioning System (GPS) sensor. Each sighting usually records an anonymized device identifier (ID), latitude and longitude coordinates, time stamps, and positioning accuracy.
Passively Collected Location Data/Mobile Device Location Data	Location sighting data generated by mobile devices, e.g., truck in-vehicle GPS sensors and cell phones.
National Truck Trip Roster	The trip roster generated from the truck sighting data.
National Truck OD Data Product	The weighted number of trips for each OD pair representing the population truck travel within and between the zones by trip distance based on the national all-trip roster.
Dwell Time	The duration of a truck staying at certain location.
Provider ID	A raw data provider's unique identifier before data aggregation
Provider Type	An indicator representing the provider type such as consumer, fleet, etc.
Data Preprocessing	Data cleaning steps including removal of sightings with invalid data entries, removal of duplicate sightings, removal of data oscillations, etc.
Truck Trip Identification	The methodology to identify truck trips from the sighting data.
Unlinked Trip	The basic unit of analysis for trips.
Trip Chain/Chained Trip	A sequence of trips with intermediate stops for non-delivery purposes.
Tracking Frequency	The number of sightings per minute.
Scalable Map Matching and Routing	The methodology to snap the sighting data to the road network and estimate the path using a routing algorithm.
Trip Distance Calculation	The methodology to estimate the trip distance for each identified trip using the sighting data and road network.
Truck Sensor Counts	The truck traffic volume collected from traffic counting programs by the truck count sensors.
Origin-Destination-Path	A sequence of truck count sensor locations that a truck trip passes in time order between a certain OD pair.