ABSTRACT

Title of Dissertation:	NATIONAL ORIGIN-DESTINATION TRUCK FLOW ESTIMATION USING PASSIVE GPS DATA
	Qianqian Sun, Doctor of Philosophy, 2023
Dissertation directed by:	Professor Paul Schonfeld, Department of Civil and Environmental Engineering

Truck travel estimation plays an essential role in the transportation field. Nationwide truck flows are particularly important for capturing long-distance truck travels. For the estimation at such a scale, the traditional way of conducting surveys is very costly and cumbersome. Nowadays, GPS data are getting popular for supporting transportation studies, with advantages of freshness, cost-effectiveness, real-world representation, high spatial-temporal coverage and resolution. Hence, utilizing GPS data as an alternative data source is worth investigating. This study proposes a comprehensive framework for achieving large-scale truck flow estimation from passive GPS data, with the United States as a study case. This study enriches the research on GPS-based travel estimation and particularly achieves the estimation at a scale as large as the United States for the first time using GPS data. The framework begins with thorough data preparation, in which an enhanced algorithm is designed for removing data oscillations. Then, truck type classification by weight class is conducted through a random forest (RF) algorithm, which enriches GPS-based vehicle classification research. The estimation is by truck type, which provides unique travel

patterns by truck type. Then, a comparative trip identification by truck type is conducted and the algorithm's robustness for such identification is investigated. Finally, an innovative weighting algorithm that integrates reinforcement learning and iterative origin-destination matrix estimation (ODME) is designed to weight the sample truck traffic according to the U.S. truck traffic population level and to mitigate the spatial bias of sample GPS data. Nationwide truck flow estimation is achieved. The results' reasonableness is discussed from multiple aspects, such as ODME accuracy, spatiotemporal biases, distance distribution, OD distribution, vehicle miles traveled, and interstate OD pairs from selected states. The products obtained from the framework are useful for many transportation studies, such as planning and operation, safety, transportation and environment, and policies. The framework not only enables large-scale truck flow estimation but also yields good accuracy and does not require excessive computation cost. It is straightforward and has a high generalizability for studies of various scales and areas. It should be widely applicable for serving transportation research and practice needs.

NATIONAL ORIGIN-DESTINATION TRUCK FLOW ESTIMATION USING PASSIVE GPS DATA

by

Qianqian Sun

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2023

Advisory Committee:
Professor Paul Schonfeld, Chair
Professor Ali Haghani
Assistant Professor Chenfeng Xiong
Associate Professor Behtash Babadi
Associate Professor Vanessa Frias-Martinez

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1. Background

The trucking industry contributes much towards economic growth. It acts as the backbone of domestic freight transportation with the highest share of freight flow tonnage among various modes including roadway, railway, airway, waterway, and pipelines. The Gross Domestic Product of transportation industry was 505 billion U.S. dollars in 2020, in which truck transportation had the highest share of 34% (Bureau of Economic Analysis, 2023). According to Freight Analysis Framework 5 (U.S. Department of Transportation, 2022), freight tonnage by truck mode is projected to grow by 42% by year 2045 from year 2020, which places trucking among the top three modes with the highest growth rate. In the development of the trucking industry, truck travel estimation plays an essential role in guiding transportation planning and operation. Travel demand estimation analyzes trip generation: trip production (trip volume produced by an origin zone), trip attraction (trip volume attracted to a destination zone), and trip distribution (distribution of trips across origin-destination pairs - OD flow). An OD matrix is a major product of travel demand estimation. In an OD matrix, columns and rows indicate the origin and destination zones respectively, and each cell shows the estimated number of trips from an origin to a destination.

Truck travel estimation contributes to a reasonable arrangement and an efficient utilization of transportation infrastructure, land-use resources, and services, such as highways, weigh and inspection stations, gas stations, rest areas, and parking lots. The design, performance, operation, and maintenance of a highway system is closely related

to truck traffic flow. This interconnection includes various aspects: efficiency, economy, safety, and sustainability. As a result of special vehicular parameters (weights and dimensions) and operating features (braking distance, acceleration, etc.), heavy vehicles such as trucks require specific geometry and capacity design of roadways (Harwood 2003). Compared to passenger cars, trucks affect highway congestion more (Sarvi 2011; Kong et al. 2016), cause greater damage to roadways (Jacob et al. 2020), and their presence increases risks of traffic accidents (Daniel et al. 2002; Freire et al. 2021). From the environmental perspective, trucks cause a higher noise level than passenger cars (Lopatin 2020). They are one of the major sources of fuel consumption (Sharpe and Muncrief, 2015) and greenhouse gases, especially nitrogen oxides (Abdull et al. 2020; Ross et al. 2011). On truck-dominated highways, several operation and management strategies, e.g., truck-only lanes, truck-climbing lanes, managed-lane facilities, and variable speed limits, are implemented in practice to improve highway systems. Besides highways, other facilities, e.g., rest areas and parking lots, whose demand is highly associated with truck flows (Vital 2021), should have sufficient availability and good accessibility for truck drivers. Under limited budgets, the investment in truck facilities or truck flow control strategies should be carefully evaluated. A reliable truck travel estimate is critically important to provide correct guidance in practice.

Additionally, transportation is an integral part of the logistics network with its effects on cost, efficiency, and service quality (Tseng et al., 2005). In traditional freight modeling, truck OD flow is a necessary component to be assigned to highway network. Common freight modeling methods, such as factored truck trip tables, commodity-

based freight models, and three-step truck models, all necessitate truck trip OD table as model inputs (Fischer et al., 2005). Total logistic cost consists of three major sections: inventory-carrying cost, transportation cost, and logistic administration cost, among which transportation cost has the largest fraction. Especially, among all subcategories of transportation cost by different transportation modes such as rail, air, water, and pipelines, trucks account for the highest transportation cost (Advanced Solutions International, Inc., 2020, 2021, 2022). In freight movement, trucking has the highest fraction of freight ton-miles, i.e. 46% in 2020 (Bureau of Transportation Statistics, 2021). Moreover, truck travel is expected to keep growing over the next two decades. According to a vehicle miles traveled (VMT) forecast study, there will be 57% more combination truck VMT and 101% more single-unit truck VMT by 2049 in the United States (Federal Highway Administration, 2022). Nationwide truck travel estimation and flow pattern analysis is indispensable for guiding and supporting national economic development.

For the above reasons, truck travel demand estimation is important for improving transportation systems, enhancing freight and logistics planning, and promoting a nation's economic growth. State and local agencies or practitioners have been implementing various methods for conducting truck travel demand estimation while there have been a data gap for nationwide truck travel demand estimation. In addition to the aforementioned reasons for the importance of truck travel demand estimation, nationwide estimation is particularly important for capturing long-distance truck travel across states. It provides a more complete picture from the perspective of the whole nation. The planning and operation of each state, instead of being fully

independent from other states, can coordinate with each other to be more efficient and obtain more benefit. For example, long-distance truck travel usually relies on interstate highway system to ensure the efficiency. With the increasing truck shipping demand, it is an essential task to enhance the current interstate system or to plan on building new interstate highways in future. Heavy truck traffic is an important factor for highway congestion. For the mitigation of severe congestion on highways, traffic obtained from the sensors is not sufficient for resolving the issue. Knowing the OD pairs generating the congestion provides insights into the real causes. Since heavy trucks often travel over long distances, statewide analysis may not be able to capture these interstate travels involving two or more states. The planning or operation procedures have to be based on a joint analysis of travel demand involving two or more states. Transferring a fraction of long-distance truck travel demand to other modes such as railway is an option for relieving long-distance shipping needs. The building of interstate highways or railways is a long-term and costly project. Its demand should be carefully examined. Truck travel demand estimation provides an essential information basis for supporting such planning and potential coordination with other modes. Nationwide truck travel demand estimation also helps improve large-scale logistic systems, especially systems that involve multiple states or regions. Both the design and the operation of such large logistic systems rely on a thorough understanding of the demand and supply.

The Commodity Flow Survey (CFS) provides OD commodity values and weights across the nation at five years intervals. However, truck flows and commodity flows are different. The former is vehicle-based, while the latter is cargo-based. They provide different insights into the transportation world. The physical features of cargo

influence truck travel. Given a specific vehicle model, moving the same tonnage of two types of cargo (e.g., iron and cotton) from an origin to a destination may yield very different truck traffic volumes. For transporting the same cargo, different truck types can also lead to different truck traffic volume. Some cargo necessitates specific truck models for safety and economy consideration. Cement requires trucks with a higher weight class such as combination trucks while small single-unit trucks are not capable of shipping it. From the perspective of logistic optimization, transportation cost is a basic component. Small-size, low-weight, and low-value cargo is relatively inexpensive to transport while large-size, high-weight, and high-value cargo requires high transportation cost. The loading status also leads to the difference between cargo flow and truck travel flow. Less loaded and empty trucks present different flows than the related cargo flows. Thus, although the CFS commodity flow survey is a valuable nationwide logistics dataset, it cannot substitute for nationwide truck travel flow. Neither a nationwide truck travel survey nor an indirect substitutable survey exists for supporting nationwide truck travel demand estimation. It is a research gap that has existed for years, which is resolved by the framework developed in this study.

For travel demand estimation, travel surveys are very commonly used. However, relying on surveys to fill this nationwide data gap necessitates a substantial investment in time, human effort, and resources. Although there have been some travel surveys, they aim to collect personal travel patterns while truck-specific travel diaries are scarce. Conducting a travel survey is time-consuming, which makes it difficult to acquire up-to-date survey data since the update period is usually one or five years. Moreover, travel surveys have some shortcomings such as low response rate (Allen et

al., 2014), underreporting trips (Bricka et al., 2012), and misreporting trips (Hossan et al., 2018). Due to limited funding, some surveys are only conducted in selected time periods and regions, i.e., with limited spatiotemporal coverage or resolution. Such surveys are unable to capture day of week, monthly, seasonal, and annual patterns and unable to serve geographical analysis at a high resolution. However, both temporal and spatial patterns are important for truck travel analysis. For instance, day of week and season significantly influence truck volumes (Yuksel et al., 2020). With a focus on normal travels, a traditional survey is not capable to support special event analysis (Kuppam et al. 2013). Instead of directly deriving OD flow from travel surveys, some truck travel models have been proposed. However, meeting the data need of these models has also been a challenge for a long time (Demissie & Kattan, 2022; Comi et al., 2013). Traditionally, travel demand analysis is conducted through an approach known as the four-step travel demand model. It mainly includes trip generation, trip distribution, and trip assignment. This traditional model may require other types of surveys and the traffic assignment on a road network is a complicated optimization problem. Neither nationwide surveys nor a feasible traffic assignment model on a nationwide road network exists. For all above reasons, an alternative way of filling this data gap is needed.

Recently, truck GPS data have been actively utilized in various transportation studies, such as travel time and delay estimation, freight performance measures, and truck parking analysis. Truck GPS data have many merits in characterizing truck travel patterns in substituting for travel survey or survey-based models. The aforementioned drawbacks of surveys can be avoided with GPS data. First, GPS data are less expensive

than in previous years. There have been increasing studies developed based on GPS data. Compared to traditional travel surveys, GPS data are more cost-effective for monitoring truck movements (IBI, 2009). Additionally, large-scale truck GPS data usually have several advantages, such as high location accuracy, high spatiotemporal resolution, and high spatiotemporal coverage. With automatic technology, the collection of GPS data is efficient. It requires less human involvement than surveys. Meanwhile, it also depends less on human efforts, so that less human error is involved. The stable and mature GPS technology avoids the uncertainty from human side. The streaming GPS data collection provides a timely and up-to-date dataset. With sufficient spatiotemporal resolution, it can flexibly characterize truck travel features in different time periods and at various geographical levels. Given a sufficient scale of GPS data, truck travel analysis across the entire nation, at state, county, or even travel analysis zone level, becomes achievable. In addition to spatiotemporal patterns, the analysis of special events, such as COVID-19 outbreak (Sun et al., 2022; Lee et al., 2020; Zhang et al., 2021), hurricanes, and sports events can also be supported by GPS data.

In spite of all these benefits, GPS data also have drawbacks. Using GPS data to estimate truck travel demand is challenging from different aspects. The major reason is that currently available large-scale GPS data in the industry are passively collected. Unlike a well-designed travel survey with detailed trip information from interviewees such as trip origin, destination, and trip distance, passive GPS data are just meaningless points without any trip level information and therefore are not ready-to-use for travel analysis. In addition, passive GPS data have obvious drawbacks such as inconsistent GPS logging frequency and missing observations, which bring difficulties to the

process of recovering trip-level information. In the current literature, a complete framework for achieving large-scale truck OD flow estimation as large as nationwide estimation in the United States from raw GPS points data cannot be found. Because of challenges or issues along with the passive GPS data, travel demand estimation requires many steps. Briefly, the framework proposed in this study resolves these problems as follows. First, raw GPS data may be noisy with data oscillations. Data oscillations, if not identified and removed, can be wrongly identified as the intermediate points of a trip leading to unreasonable travel statistics, such as very long travel distances. Missing observations is a common drawback of passively-collected GPS data. With limited observations of a truck, the derived trips become more sensitive to the existence of data oscillations. Especially, nationwide truck travel demand estimation contains a significant portion of long-distance travel. Data oscillations bring about misleading travel patterns. For example, a local delivery truck with data oscillations may show longer-than-normal travel distance and brings noises to vehicle type classification, which may finally be misidentified as a heavy truck. The data oscillations are carefully examined and removed in this study. Second, different truck types present different driving behaviors and spatial-temporal mobility features. For instance, local delivery trucks usually conduct shorter and more frequency trips while heavy trucks may travel for hours and produce longer trips. For algorithms driven by GPS data, it is important to measure their robustness in different scenarios by truck type. For such algorithms that are sensitive to truck types, parameter values should be carefully tested and decided. Generally, passive GPS data usually do not contain truck type information. This study fills up the missing truck type information, investigates the robustness of

algorithms from aspect of truck type, and finally conduct truck travel demand estimation for each truck type. Third, trip is the basic analysis unit for travel analysis, which is usually missing from the passive GPS data. To approximate the real-world truck travel patterns and activities, raw GPS points need to be converted to meaningful trips. There have been numerous trip identification algorithms driven by passive GPS data, which are actively applied and perform well in various studies. This study does not intend to propose a different trip identification algorithm. Instead, a prevailing type of GPS-driven trip identification algorithm is applied. Most importantly, the algorithm's sensitivity to truck type is comprehensively examined, which is a major contribution to this field. Fourth, currently available truck GPS data in the industry provide a sample from the population. No matter how large the sample data size is, a weighting procedure needs to be conducted to produce estimates for the population. Otherwise, the derived statistics are not representative of the U.S truck population. Weighting has been a challenge for many reasons. Existing weighting methods are not applicable to nationwide scenario due to computation complexity of nationwide highway network or unavailability of necessary data inputs such as penetration rate that is usually missing, unknown, or not provided by data providers. Besides, nationwide available ground truth data are very limited. Fifth, passive GPS data collection does not follow a strictly-designed sampling procedure, so the sample bias probably exists, which should be resolved. Common biases in GPS data are temporal bias and spatial bias. Temporal bias mainly results from missing observations of the GPS data. Due to the backend technology limitations, it is normal that some GPS observations are missing for a while or even for hours. For example, when the GPS device is out of

battery or under a long tunnel, the receptor fails to accept the GPS signal. The spatial distribution bias mainly comes from the different hauling service areas of the companies or organizations providing the GPS data. Since the sample GPS data only come from a subset of companies or organizations, the measurement of spatial bias is of great importance for deriving reasonable travel demand estimations or statistics. Additionally, truck type bias is the third type of bias that must be considered for the topic in this study. Trips generated from different truck types have obviously different features regarding trip production, trip attraction, and trip distribution. Local delivery trucks and long-haul heavy trucks vary greatly regarding service area, hauling distance, and on-duty time window, and hence produce very different OD flow patterns. Therefore, three types of biases in total are considered in this study.

In sum, the significance of conducting nationwide truck travel demand, the existing data gap of national truck OD flow, and the benefits of utilizing GPS data to fill this data gap initiate and motivate this study. All of the challenges or issues mentioned above are resolved by the proposed framework and finally a large-scale truck OD flow matrix by truck type for the entire United States is derived from passively-collected truck GPS data.

1.2. Objectives

Nationwide OD truck travel flow data play a very important role in a variety of transportation studies and practices. However, currently such data do not exist. In comparison to the traditional way of conducting travel survey to fill up this data gap, GPS data have shown many advantages. Such a framework that derives nationwide

truck travel flows from GPS data does not exist. In summary, this study has the following objectives.

(1) To develop a feasible framework that enables nationwide truck flow estimation from passive truck GPS data. Currently existing frameworks are either not applicable to the research need of this study or not feasible for such a large scale as nationwide case. This study aims to successfully derive nationwide truck flows from passive GPS data, which is the major objective. Moreover, to ensure the high feasibility in practice, the data utilized are either open-source or prevailing in this research field. The implementation does not just consider the accuracy but also tries to reduce computation cost and improve efficiency.

(2) To develop a complete framework that incorporates all fundamental parts to fill research gaps. As previously discussed, several fundamental research tasks are needed to derive truck flows from GPS data, along with which research gap, especially nationwide weighting method, exists. Based on current literature, either improvements are made to existing methods or innovative methods are developed if research gap exists.

(3) To develop a thorough framework that addresses the previously mentioned challenges, such as mitigating sample biases, large-scale weighting, and high computation cost.

(4) To develop a practical framework that is computationally efficient and yields desirable accuracy level. The algorithms or methods involved with the framework should not require excessive computation cost. Computation cost may become a bottleneck for the application of large-scale GPS data in practice since some well-defined algorithms cannot ensure both high accuracy

and low computation cost. This study aims to achieve a desirable accuracy level without much compromise in computation cost.

(5) To develop a high-generalizability framework that is applicable to various spatiotemporal scales. GPS data are advantageous for high spatiotemporal resolution. To make the best use of GPS data, a flexible framework that can be generalized to studies of different spatiotemporal scales is preferred. The framework can be applied to but is not limited to a nationwide scenario. It can also be implemented for statewide or regional studies.

(6) To provide an empirical reference on nationwide truck flow patterns from GPS data. These are very valuable information.

1.3. Contributions

The major contribution of this study is the achievement of large-scale OD truck flow estimation with United States as the case study. To the knowledge of the author, no such study has ever been done at a similar scale. The derived OD truck flow matrix is a valuable data product that can support studies in many fields. The proposed framework not only overcomes the current research gap but also addresses challenges and issues along with the application of passive GPS data to truck travel estimation. Specifically speaking, this paper has the following contributions.

(1) A complete framework of estimating truck flows based on passive GPS data is specifically designed for large-scale studies, making it feasible to achieve nationwide truck OD estimation in the United States for the first time. This study fills the data gap of nationwide truck flows and fills the research gap of large-scale (e.g. nationwide) truck flow estimation from GPS data. The

framework yields good accuracy without heavy computation burden. It has a high generalizability, flexibility, and feasibility for studies in different regions and scales. It is straightforward and should be widely applied in practice.

(2) The derived OD matrix is an essential data product that is needed in many fields. The derived OD matrix directly reflects the truck traffic flow for each OD pair, especially for these long-distance interstate OD pairs. In addition to the OD matrix, some other products can be derived with the developed framework. For example, the weighted truck trips can be projected to the highway network, so that truck travel flows on corridors can be obtained and bottleneck areas can be identified. Some statistics such as vehicle miles traveled and average daily truck traffic are also obtainable.

(3) With the United States as the case study, the derived result provides insights into the characteristics of nationwide OD truck flow from several aspects. These are the unique and empirical references from passive GPS data. In addition, the estimation is conducted by vehicle type: (a) light and medium weight trucks; (b) heavy trucks. The derived results shed lights into the different patterns between light-medium and heavy trucks.

(4) An elaborated pattern-based method for data oscillation identification is designed, which improves the existing pattern-based methods by considering more abnormal moving patterns and reducing location uncertainty in an efficient way.

(5) This study considers three types of biases in passive GPS data for fulfilling truck flow estimation - spatial bias, temporal bias, and vehicle type bias. This is critical for ensuring the representativeness of result.

(6) Vehicle type classification is conducted on the raw GPS points through a random forest (RF) algorithm. Current vehicle type classification methods are mainly designed based on traditional data sources (e.g. radar, loop, and video) and GPS-data-based methods are very limited. Hence, this paper enriches this study field. To the knowledge of the author, this is also the first study that conducts RF-based vehicle classification from GPS data. A comprehensive set of input features is explored, among which some new features are innovatively designed and prove to be significant for differentiating vehicle types through an RF algorithm.

(7) Truck trip identification is conducted by truck type. A popular type of trip identification methods is implemented and, specifically, its application to different truck types is discussed. Many studies usually arbitrarily decide the threshold values for trip identification from GPS data without investigation of truck type. This study for the first time investigates the algorithm's sensitivity to the truck types, which provides a comparative reference.

(8) An iterative reinforcement-learning-based Origin Destination Matrix Estimation (ODME) method is innovatively designed, which weights the sample truck flows derived from GPS data to the U.S. truck traffic population level. The method achieves a desirable accuracy level. This fills the research gap of achieving large-scale ODME from GPS data. The largest scale of

current ODME methods is state-level such as for Florida and Indiana. The model complexity and dense national road network makes the traditional ODME method infeasible for national application. This study provides an alternative way of matching with the sensors on the highway network.

1.4. Outline

The remainder of this study is organized as follows. Chapter 2 provides a comprehensive literature review. Chapter 3 describes the data used for this study and preprocessing work. Chapter 4 introduces the heuristics algorithms designed for the identification and removal of data oscillations. Chapter 5 presents the classification of raw GPS point data into two truck types: light-medium weight trucks and heavy weight trucks. Chapter 6 describes the preparation of truck trip profile and specifically explores the trip identification algorithm's robustness in the application of different truck types. Chapter 7 proposes an innovative weighting algorithm with the integration of reinforcement learning and ODME process. Chapter 8 presents major results and discusses the reasonableness and validation of derived results. Chapter 9 summarizes the whole study and highlights the practical significance of the developed framework. Chapter 10 discusses future work.

Chapter 2: Literature Review

2.1. Research and Practices Based on Truck GPS Data

With the development of processing GPS data and extracting travel statistics from it, more and more studies in truck transportation field have been making the best of truck GPS data for various research and studies. Some studies utilize truck GPS data to identify trip ends and construct the truck trip profile (Thakur et al., 2015; Aziz et al., 2016). Along with the trip identification studies, some trip chaining methods are developed to complete the trip profile based on different application needs. Instead of starting from trip identification, some studies conduct tour identification analysis without the need of chaining the trip segments. Both trip identification and tour identification studies can serve as the basic of truck travel demand estimation or prediction. For example, some studies develop trip- or tour- based models or frameworks for supporting travel demand analysis (Kuppam et al., 2014; You and Ritchie, 2019; Demissie and Kattan, 2022). Besides, some research estimates or forecasts the travel time, travel delay, and the reliability of travel time for selected intersections or road segments using the GPS data. Roadway congestion and bottleneck analysis is also a major application of the GPS data. Other use cases of the GPS data include truck parking analysis, freight activity analysis, truck route choice analysis, freight simulation and modeling, policy and investment analysis.

Government agencies also conducted various projects with regards to truck GPS data. The Florida Department of Transportation (FDOT) did an analysis of freight performance measurement, modeling, and planning by utilizing truck GPS data early

in 2010. It derived a database of truck trips with origin and destination Transportation Analysis Zone (TAZ) based on Florida statewide model. Florida Department of Transportation (FDOT) Transportation Data and Analytics Office (TDA), in coordination with the Freight and Multimodal Operations Office (FMO) and District Freight Coordinators (DFCs), conducted a statewide study to analyze parking supply and utilization. Their study identified critical truck parking needs and proposed corresponding solutions using truck GPS records from the American Transportation Research Institute in 2019. FDOT in 2017 conducted a truck route choice modeling analysis using the truck GPS data, which measures the diversity of travel paths between origin and destination pairs in metropolitan regions of Florida and assesses the algorithms of route choice set generation. The Minnesota Department of Transportation developed a methodology for analyzing the performance of heavy commercial trucks with respect to freight mobility and reliability in the twin cities metro area using the truck GPS data in 2014. The Arkansas Department of Transportation conducted a truck activity analysis in 2019 for the need of statewide freight modeling and planning, which shows several potential usages of truck GPS data – truck parking utilization patterns, travel time delays, and impacts of waterway ports. The National Center for Freight and Infrastructure Research and Education developed Freight Performance Measures (FPMs) to meet Moving Ahead for Progress in the 21st Century Act (MAP-21) objectives and to apply the methodology in the National Center for Freight and Infrastructure Research and Education (CFIRE) region by utilizing the truck GPS data in 2016. A case study was conducted for the state of Tennessee. The Puget Sound Regional Council (PSRC), Washington State Department of Transportation (WSDOT),

and the University of Washington (UW) cooperated on collecting and making use of the truck GPS data from commercial, in-vehicle, fleet management systems in 2011.

2.2. Truck Travel Demand Estimation Using GPS Data

A complete framework for estimating truck OD flows from passive truck GPS data specifically for a large-scale case, such as the entire United States, cannot be found in the literature. There are many studies at smaller scales. Kuppam et al., 2014 make some first attempts at developing a tour-based truck travel demand model from ATRI GPS data with a case study in Phoenix. It focuses more on the tour generation and stop generation along the tours. Expanding the sample travel demand to the population level is missing from their work. Bernardin et al., 2011, produce an Indiana statewide truck trip OD table from the GPS data using an ODME process to match the truck sensor counts. Zanjani et al., 2015 achieves a statewide truck travel estimation from ATRI GPS data. It mainly focuses on an ODME optimization algorithm to weight and expand the sample trips extracted from GPS data to match with the observed truck counts on some road segments. You & Ritchie, 2019, build an optimization model with entropy maximization to estimate drayage truck demand at the San Pedro Bay Ports (SPBPs) complex in Southern California with GPS data as the data inputs. The entropy maximization is a non-linear problem (NP) with linear constraints. The number of constraints increase as the case study size increases with more and more OD pairs. This limits its application to a small-scale study. All three studies (Bernardin et al., 2011; Zanjani et al., 2015; You & Ritchie, 2019) build an optimization model, which is too computationally complicated to be applied to a large-scale study such as nationwide network. There are some other studies that propose the combination of truck GPS data

and the survey data to analyze the truck movements (Laranjeiro et al., 2019; Demissie and Kattan, 2022). Demissie and Kattan, 2022, conduct another truck travel estimation for the province of Alberta, Canada. They uses a multinomial logit model to estimate the trip distribution using both GPS data and the outputs of a provincial travel demand model relying on traditional surveys. Its major contribution is its combination of passively collected truck GPS data and actively collected survey data to derive truck travel demand. However, their study is based on the distribution directly obtained from the sample GPS data. Sample bias or weighting is not discussed.

2.3. Other Related Studies

Although complete frameworks for large-scale OD truck travel flow from passive GPS data are still underdeveloped, there have been studies focusing on some fundamental parts. A complete data cleaning and assessment is a prerequisite to ensure the data quality. The major task is to identify and remove data oscillations. The second essential part is vehicle type identification from GPS data. The third part is recovering meaningful trips from meaningless raw GPS points. The last basic part is weighting and expanding since the GPS data is just a sample that cannot reflect the population-level statistics. There are challenges, issues, or research gaps with each of these parts, which are discussed in the following sections 2.3.1 – 2.3.4. The pros and cons of current methods are also summarized.

2.3.1. Data Oscillation Detection

Large stream GPS data are getting popular in many fields. To ensure reliable inferences from them, a complete data preprocessing is essential. Otherwise, the

mobility inferences from GPS data might be undermined or biased (Wu et al., 2014; Wang & Chen, 2018). Studies on data preparation are very few. Based on literature review, a major cleaning work that GPS trajectory data particularly needs is the detection and removal of data oscillations, also called outliers or data jumps. Most outlier detection methods are statistical tests based on the statistical distribution (e.g. normal distribution, gamma distribution, etc.) of a single variable (Hawkins, 1980; Barnett & Lewis, 1994; Knorr et al, 2000). This method does not work for the passively collected GPS data, which usually do not follow a specific distribution. Current outlier detection methods for trajectory data or GPS data can be classified as speed-based, distance-based, partition-and-detect, density-based, clustering-based, classification, and pattern-based methods.

A speed-based method simply removes points with an extreme high speed from its previous point. Several different speed thresholds are used in studies, such as 200 km/hr and 200 mi/hr (Thiagarajan et al., 2009). This type of method is too simple to be reliable. In reality, the cases of data oscillations in GPS data can be much more complicated. For a simple example, a point with an extremely high speed from its previous point may actually be a normal point. However, this method will remove this point and keep removing its following points too. Many false positive cases will be caused by this method.

Knorr et al, 2000 proposes a distance-based outlier detection method, which can detect outliers from trajectory data. It first characterizes each trajectory by a multi-dimension vector with a set of mobility features (i.e. start and end point, number of

points, heading, and velocity). Then a distance function is built by the weighted sum of the difference between the vectors of two trajectories to measure the similarities between paths. This method has some limitations. If there is just a small segment showing abnormal features in a very long trajectory, this abnormality is probably averaged out across the entire trajectory so that cannot be identified (Lee, et al., 2008). To make outliers stand out, most of the trajectories in a dataset have to share similar mobility behavior. However, in practice, trajectories in the raw GPS data show various mobility behavior even if they are not outliers. Lastly, since this method removes the entire abnormal trajectory, there may be unnecessary data loss.

Instead of identifying and removing abnormal trajectories, some studies focus on the outlier segments in a trajectory. These studies apply a partition-and-detect model (Lee, et al., 2008; Yang and Tang, 2016; Zhang and Wang, 2011; Krishnan, et al., 2017; Yu, et al., 2014, Gupta, et al., 2014). Lee, et al., 2008 firstly propose the partition-and-detect model to first partition trajectories and then remove abnormal sub-trajectories. They design an approach to detect abnormal sub-trajectories from aspects of position and angle by taking neighboring trajectories as baselines. A major limitation of this method is that the detection of outliers is dependent on neighboring trajectories. With the absence of neighboring trajectories, the standalone sub-trajectories are determined to be outliers and therefore are removed. This method requires a high sampling rate from spatial distribution point of view. However, for a nationwide truck GPS dataset, this requirement is usually cannot be satisfied and the data loss is excessive in this case.

There are also density-based outlier detection methods. Breunig et al., 2000 first propose this idea by defining local outlier factor (LOF), which measures how a point is isolated from its neighboring points. Points with a high LOF are determined to be outliers. Neighborhood size is very critical for this method since if it is very small outliers may not be identified and if it is very large much computation cost is required. The parameter for defining the neighborhood size turns out to be very sensitive (Papadimitriou et al., 2003). This method does not fit the research need in this study. The LOF involves too much computation between its neighboring points, which is not a good choice for a nationwide GPS dataset. Passively collected GPS data may have low frequency for a while making it normal to have standalone points that are geographically distant from others. They are not outliers and may become an intermediate point on a trip, but will be identified as outliers by density-based method.

Similarly to density-based methods, some clustering-based methods (Ester et al., 1996; Guha et al., 1998) can detect outliers based on the spatial densities. However, clustering-based methods are originally intended to produce clusters instead of to identify outliers. The points that cannot form or be involved in a cluster are identified as outliers, which may actually be normal points in other research contexts. Li et al., 2007 propose a classification-based method for detecting abnormal trajectories. However, a good training dataset is usually unavailable in practice.

One more type of method for detecting data oscillations is pattern-based methods. Lee & Hou, 2006 propose a pattern-based method for identifying the ping-pong transitions in cellular data (the cell that a mobile connects to changes in a short

time making quickly its geographical location changes back and forth). This method defines two movement patterns of data oscillations by the sequence of cellular towers. Lovan et al., 2013 consider spatial temporal features such as speed in addition to the movement patterns to identify ping-pong phenomenon. Wang & Chen, 2018 propose a method that first builds clusters from cellular data (triangularized latitude and longitude data) to reduce location uncertainty and then identify data oscillations by defining a set of movement patterns. Wu et al., 2014 identify data oscillations in cellular data by defining the movement patterns based on two major ideas: stable period and impossible moving speed. Four heuristics are proposed to define abnormal movement patterns with spatial-temporal features including speed, distance, and time. All these methods are designed specifically for the data oscillations (ping-pong phenomenon) from cellular tower data or cellular data. Since GPS data show more granular and complicated data oscillations, these methods are not sufficient. Despite this, the author sees the potential of extending such pattern-based method to handle the data oscillations in GPS data enlightened by the ideas of reducing location uncertainty, defining movement patterns by location sequence, stable time period, and considering the spatial-temporal features. So far, this pattern-based method has the greatest flexibility and adaptability to the truck GPS dataset used in this research.

2.3.2. Vehicle Type Classification

Vehicle classification methods have been investigated in many studies (Coifman & Kim, 2009; Eikvil et al., 2009; Kafai and Bhanu, 2012; Ma and Grimson, 2005). A major factor that differentiates these methods is the input data source. Common sources are sensors (e.g. pneumatic tubes, loop detectors, weight sensors,

radar technology, infrared and acoustic devices), satellite images, and surveillance cameras. Each data type uniquely determines one group of vehicle type identification or classification methods. In other words, each group of methods is not-ready-to-use for other data types meaning a low generalizability. Moreover, these methods may yield errors or show limitations in some cases as comprehensively discussed by Sun and Ban, 2013. For instance, congestion highly affects the classification when using loop sensors. The study scale might be limited due to insufficient application of aforementioned technologies. When it comes to large areas, it may be too costly to apply these traditional methods (Sun and Ban, 2013).

With the increasing interest in GPS data, there are a few studies starting to classify vehicles through GPS data (Sun and Ban, 2013; Simoncini et al., 2016; Simoncini et al., 2018; Dabiri et al., 2020). Overall, these studies first define a set of features and then implement a machine learning method as the classifier. Sun and Ban, 2013 uses a support vector machine (SVM) to classify trucks and passenger cars using a small (136 vehicles) and high frequency (at 1 second or 3 seconds intervals) active GPS sample from experiments. Since in reality such high-frequency data are usually not widely available, Simoncini et al., 2016, claim, for the first time, to address the problem of vehicle classification from low-frequency GPS data. They conduct a binary classification (i.e. light- and heavy- duty truck) based on an active GPS sample (2000 vehicles) with a frequency at 90s or 120s interval. Although the same SVM algorithm is applied, they propose some new features based on speed, distance and acceleration, which show better performance in low-frequency scenario. Simoncini et al., 2018 further improve the performance of vehicle classification (i.e. light-, mid-, and heavy-

duty truck) from low-frequency GPS data through a deep neural network method, which is applied to a large active GPS dataset (96,338 vehicles) with an average 90s interval. This deep learning method produces a more accurate result than RF and SVM. Dabiri et al., 2020 propose another deep neural network method to classify vehicles by weight class based on a high frequency (at 17 seconds intervals) and large-scale (20 million trajectories for four months) GPS dataset. These two deep learning methods have two common drawbacks. One is the high computation cost, and the other is the low interpretability, since they automatically learn the data without much revelation on important features. One more study by Zanjani et al., 2015 divides the identified trips from GPS data into medium- and heavy- duty truck trips through arbitrary trip distance and trip frequency thresholds. Sun et al., 2020 derive four vehicle types using k-means clustering method based on GPS data (68,613 vehicles, at 30s to 60s intervals) from floating cars. Unlike above research, their study focuses on mobility patterns (i.e. morning-activity, long distance traveling, frequent activity, and evening activity) instead of weight classes. The major feature that differentiates the four types is the roadway usage. Although not many studies conduct GPS-based vehicle classification, some other studies help provide insights into this topic, which aim to extract driving behavior features from GPS data (Dong et al., 2016; Guo et al., 2018; Chen et al., 2019). Based on all studies discussed above, some significant features that can help differentiate vehicle types are as follows: roadway usage, moving angle, aggregated statistics (i.e. median, mean, variance, maximum, 25 quartile, and 75 quartile) of speed, acceleration, time, and distance, similar statistics over a sliding window, and trip rate. These practices also provide the correspondence between vehicle weight classes and

the thirteen vehicle categories from FHWA. Simoncini et al., 2016 and Simoncini et al., 2018 state that light-duty corresponds to class 2 and 3 and heavy-duty corresponds to class 5 and higher. Dabiri et al., 2020 apply the following correspondence based on vehicle gross weight: (1) less than 14,000 lbs ~ classes 1-3; (2) between 14,000 lbs and 26,000 lbs ~ classes 4-6; (3) larger than 26,000 lbs ~ classes 7-12.

Based on literature review, SVM and deep learning are two state-of-the-art methods for vehicle classification. Another popular machine learning classifier is RF. As traditional machine learning classifier, RF and SVM have been widely applied for various classification problems. It is usually said that RF works better for multi-class problems and SVM is better for binary problems. Yet, the two may yield nearly the same accuracy levels in some cases. For vehicle classification problem, Simoncini et al., 2018, compare the three methods regarding accuracy and computation time. Although deep learning yields the highest accuracy, RF runs much faster than deep learning and SVM. In their tests, RF takes 20 min while SVM takes at least 10 days; RF takes 10 s while deep learning takes 5 min. Since this study necessitates data processing on a large-scale dataset, computation cost becomes a critical factor. Therefore, RF makes it advantaged for being computationally low-cost. Additionally, RF has a higher interpretability than deep learning method. There have been many methods for measuring the relative importance of input features and how each input feature is interacting with the final output in an RF model. However, deep learning method automatically learns from the data, which is done in the hidden layers and cannot be explained. In sum, RF has the traits of good classification accuracy, low computation cost, and high interpretability and is therefore utilized in this study.

Moreover, SVM is set as the baseline model, which shows similar accuracy but requires much more computation time.

2.3.3. Truck Trip Identification

For many applications of passive GPS data to the transportation field, trip identification is an essential step, which is the conversion of raw GPS points into meaningful trips. So far, different methods have been proposed. Overall, there are three common types of truck trip identification methods. The first type is rule-based methods; the second type is clustering methods; the third type is machine learning methods.

This type of method identifies trip ends by considering the speed, dwell time, and distance between two consecutive GPS points. Some studies solely apply the average speed to determine truck trip ends (Ma et al., 2011; Gong et al., 2015; Zanjani et al., 2015; Gingerich et al., 2016; Sarti et al., 2017). Different speed thresholds are applied, such as 1 km/h (Gingerich et al., 2016), 5 km/h (Sarti et al., 2017), and 8km/h (Zanjani et al., 2015). These methods may identify false positive truck trip ends since the status of being low-speed does not indicate a meaningful trip end in some cases, such as waiting for a green signal, stopping due to congestion. Instead of speed, dwell time is used for identifying trip ends by some studies (Camargo et al., 2017; Gingerich et al., 2016; Laranjeiro et al., 2019; Thakur et al., 2015). It is assumed that if a device stays at somewhere for a long time, there is a trip end. The threshold of dwell time varies greatly, which mainly relates to the truck type. For example, a local delivery truck stops frequently to distribute goods and at each stop it stays for a short time. Therefore, smaller threshold values are usually applied to short-haul trucks, such as 3

minutes and 5 minutes (Camargo et al., 2017). In contrast, long-haul trucks require more stopping time for loading or unloading goods. Some studies apply a larger threshold value of dwell time, such as 10 minutes (Aziz et al., 2016), 15 minutes (Gingerich et al., 2016; Thakur et al., 2015), and 20 minutes (Laranjeiro et al., 2019). A few studies use distance to identify trip ends (Calabrese et al., 2011). Thakur et al., 2015 design an algorithm involving ten parameters including speed, time, distance, and several cumulative features such as cumulated distance and time, origin and destination dwell time to identify truck trips. In one previous work co-authored by the author (Zhang et al., 2021), a recursive trip identification algorithm is proposed, which incorporates speed, dwell time, and distance between any two continuous GPS points of a device. On the one hand, rule-based heuristics are straightforward and easy to be applied. On the other hand, the determination of threshold values depends on the features of GPS dataset and the study setting. In this study, the trip identification algorithm co-authored by the author (Zhang et al., 2021) is applied. That previous work focuses on personal trip identification from passive GPS data while in this study the algorithm is applied to the truck side for the first time. Moreover, trip identification by truck type is investigated. This present study provides a good reference on threshold values for different truck types in the same context.

The second popular type of methods for trip identification from GPS data are clustering methods. Representative clustering methods include K-means method, Density-based Spatial Clustering of Application with Noise (DBSCAN), and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). So far, many variations of density-based clustering methods have been

applied to identifying trip ends, such as revised DBSCAN (Palma et al., 2008; Tran et al., 2011), T-DBSCAN (Chen et al., 2014), C-DBSCAN (Gong et al., 2015), and DBSCAN-TE (Gong et al., 2018). K-means method requires the pre-determination of the number of clusters. For a nationwide GPS dataset containing device with a wide range of travel distance, it is unreasonable to have such pre-determination. Density-based clustering methods have a high requirement on the GPS data quality. They require high-frequency GPS data (Gong et al., 2015). Missing observations are very common in the nationwide truck GPS data leading to a biased spatial distribution. This phenomenon greatly undermines the performance of density-based methods (Fu et al., 2016; Gong et al., 2015).

Recently, some studies apply machine-learning methods to accomplish trip identification. Yang et al., 2021 implement a random forest model to identify trip ends using cellular phone and points of interest data. Ferreira et al., 2021 implement a neural network method to identify trip ends from GPS data. There also some other studies that apply machine-learning methods to extract activities from the GPS data (McNally and McGowen, 2006; Montini et al., 2014). The major deficiency is that these machine-learning methods require a validation dataset, which is not available from the passively collected GPS data.

2.3.4. Traffic Weighting and Expansion Methods

Since GPS data is just a sample of truck movement, estimating truck travel demand from it necessitates a weighting and expansion process. Currently, just a few studies discuss how sample truck trips can be expanded to the population. On the

passenger side, several weighting techniques are proposed. A state-of-the-practice method for weighting personal trips is the penetration rate method, which utilizes the population data from the United States Census Bureau. The trip profile from the National Household Travel Survey provides insights into the population-level statistics about personal trips. Hence, Zhang et al., 2021, design a two-fold process that accomplishes the weighting at both the device- and trip-level. However, such weighting method is not applicable to the truck trips. The home location of personal trips can be retrieved by various home location identification methods so that residential sampling rate (i.e. penetration rate) can be imputed to help weight up the sample GPS trips. Since each truck GPS device is bundled with a truck, penetration rate weighting method needs the dataset of the total number of trucks operating at each area. Unfortunately, these data are unavailable. The truck GPS data providers do not have the penetration rate either.

Truck sensor counts data collected by sensors installed on highway network are the major ground truth data that can be utilized for weighting and expanding the sample GPS truck trips. Current studies of utilizing GPS data to estimate truck travel demand are usually conducted by first producing a seed OD matrix from the GPS data and then conducting an ODME process to match with the observed truck sensor counts (Bernardin et al., 2011; Zanjani et al., 2015; Bernardin et al., 2017). ODME means origin-destination matrix estimation. It is a traditional approach that has been implemented by transportation researchers and practitioners for a long time. An ODME process includes the following steps: 1. Preparing a seed OD matrix; 2. Assigning the OD matrix to the road network by traffic assignment models; 3. Adjusting the OD

matrix by minimizing the difference between assigned traffic flows and the observed traffic counts on the road links; 4. Repeating the previous two steps until the difference is reduced to a desired level. This ODME-based weighting method involves traffic assignment, which estimates the flows on the road network based on the travel cost (usually measured by travel time) estimate of alternative paths that can carry the given traffic volume. Various traffic assignments models have been developed for remapping the traffic volume. A common and essential data input of traffic assignment models is the travel cost profile, which is usually only available for small study areas. No such travel cost estimate data exists at the nationwide level. Besides, a traffic assignment algorithm is usually based on complicated optimization models. With increasing road network size, the problem gets so complicated that convergence may not be achieved. A national scale traffic assignment is undoubtedly impossible. To serve the research need in this study, a revised ODME process that resolves the traffic assignment challenges is needed.

Some other weighting methods utilizing the traffic sensor counts are discussed in the literature, including simple scaling to counts and iterative proportional fitting. Simple scaling uses one single factor derived from the difference between sample OD and observed traffic count to weight up the sample. This method is too simple to ensure the accuracy level. Iterative proportional fitting (IPF) repeatedly adjusts the sample OD matrix until the marginal totals are matched with a group of observed traffic counts at sensor locations. As additional sensor locations are involved, the problem gets increasingly complicated. For a nationwide study case, the number of sensor locations is in thousands making the IPF method computationally inapplicable.

Chapter 3: Data Description

The data used in this study are from two major truck GPS data providers in the United States, whose data have been actively used by many researchers for either static or even dynamic studies. Both datasets have a spatial coverage as high as all fifty states and Washington D.C. In a streaming format at several seconds or minutes, hour-of-day, day-of-the-week, weekly, monthly and seasonal analysis can be feasible. Both datasets are passively collected and contain the basic location information including device ID, timestamp, instantaneous speed, latitude, and longitude. The two datasets also differ in some aspects. First, dataset I additionally provides the vehicle weight class of each truck while dataset II does not. Dataset I's composition by truck type is displayed in Table 1, showing that it mainly consists of light- and medium- duty trucks and a small fraction of heavy-duty trucks. In comparison, dataset II mainly consists of heavy-duty trucks, as described by data provider. The vehicle classification in this study is for grouping dataset II into different truck types. Second, dataset I has some trucks that are not consistently observed for technology reasons, which leads the change of device ID in the data, but this is not an issue in dataset II. For solving this issue in dataset I, a probability-based trip chaining algorithm is applied, which is described in trip identification section.

Particularly for this study, one month of (2020 January) truck GPS data across the United States from the two providers are used. One-day data are displayed in Figure 1 with a zoomed-in portion in New York City. It shows that the data have a very high spatial coverage and distribution density. Temporally, dataset I devices are observed

for 3 days on average and dataset II devices are observed for 12 days on average. The maximum observed days are 31 and 29 days respectively for dataset I and II. The difference in average observed days is consistent with the domination of the two datasets by truck type. In total, there are 9 million trucks with 12 billion GPS logging points are used for this study.

Table 1. Truck type composition in dataset I.

Truck Type	Light weight	Medium weight	Heavy weight
Percentage	74.9%	24.6%	0.5%

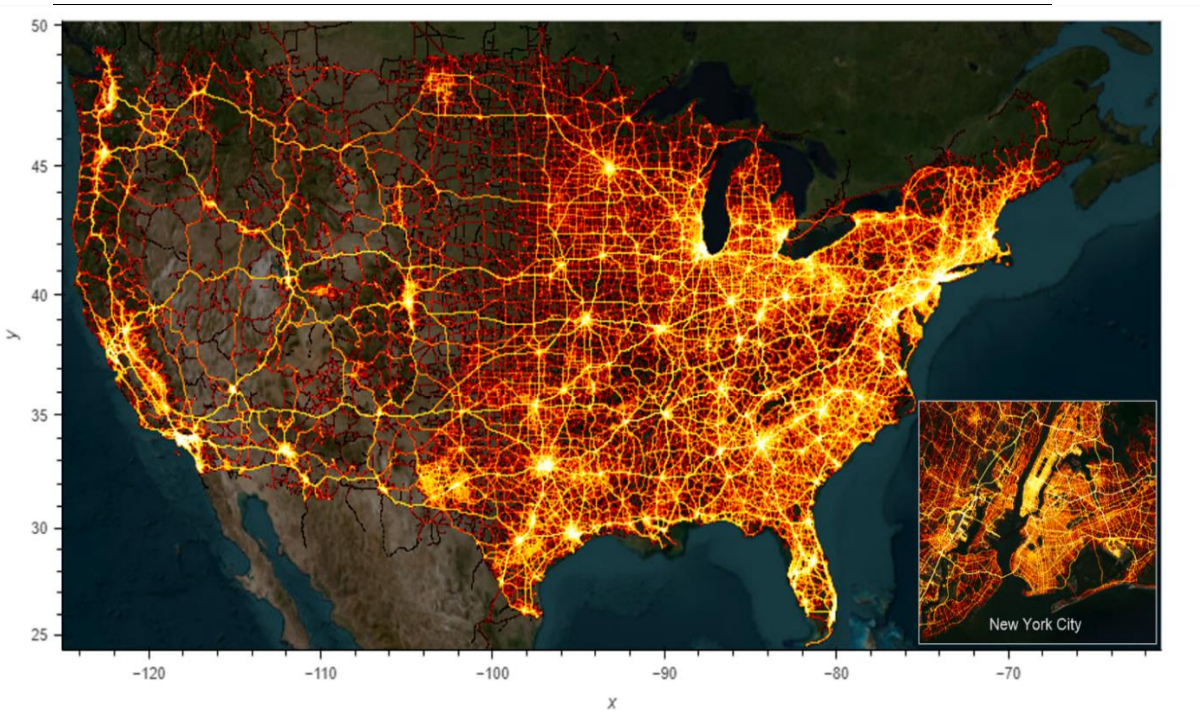


Figure 1. Example of one-day truck GPS data used for this study.

For a deeper understanding, some quality metrics are designed and summarized in Table 2. These metrics are believed to be important for a data-driven travel pattern

analysis. Both datasets are low-frequency with a median value of 1 minute interval and the distribution is biased to longer time intervals. On average, dataset I truck is observed at 4.7 minute intervals and dataset II truck is observed at 2.2 minute intervals. Being low-frequency is very common in the industry for saving costs, which makes it more practically meaningful to handle low-frequency GPS data. Although low-frequency, both datasets have a high number of average daily observations (236 and 509 for dataset I and II respectively) and a high number of average daily active hours (7 and 14 for dataset I and II respectively). Dataset II is higher because it mainly contains heavy trucks.

Table 2. Quality metrics of raw GPS data.

Quality metrics	Description	Dataset I		Dataset II	
		Mean	Median	Mean	Median
Frequency	Time interval between two consecutive sightings in seconds	287	60	134	60
Daily frequency	Average daily number of sightings by device	236	193	509	375
Active hours	Average daily number of observed hours by device	7	6	14	14

Other than GPS data, some other datasets are used for different purposes. 2020 truck count data from FHWA's Travel Monitoring Analysis System (TMAS) are used as the ground truth data for weighting the sample GPS data and mitigating the spatial bias. The 2020 Freight Analysis Framework (FAF) nationwide State-level OD commodity flow data, estimated from 2017 CFS data, are used as a major validation source. The 2020 Highway Performance Monitoring System (HPMS) road network is used for providing roadway usage regarding functional class. 2020 Smart Location Database (SLD) is used for extracting built-in environment features to characterize the

typology of truck movements and therefore to serve as an important input for vehicle classification. It is an open-source dataset. In this study, the nationwide road network is obtained from OpenStreetMap (OSM), which is more refined than that from HPMS. Both sample GPS trips and truck count sensors are map-matched to OSM network. This is the basis for deploying the proposed iterative reinforcement-learning-based ODME method. Point of interest (POI) data are used in the step of trip chaining for preparing truck trip profile.

Since the whole study is built upon the GPS dataset, the quality of the dataset should be assured before conducting any analysis. Noises, such as duplicate sightings, missing values, invalid values and especially data oscillations, are removed since they may distort the analysis. Data preprocessing aims to remove irrelevant, incorrect, and invalid data records, and to standardize the dataset with the designed format. This work is essential for resolving the errors and noises along with the raw dataset. In detail, the data preprocessing proceeds through the following steps.

- Scrubbing irrelevant data: raw GPS data may contain attributes that are irrelevant to the research need in this study, such as geospatial type, probe source type, and movement type. These attributes should be dropped.
- Dropping duplicate sightings: if multiple locations for the same timestamp exist, then the sighting with the best accuracy is kept if accuracy measure is available.
- Removing missing values: records with null values should be removed.

- Removing invalid sightings: the value range of each attribute is imputed and the records with values out of the reasonable value range are removed. For example, GPS sightings with latitude and longitude equal to zero should be eliminated. The UTC timestamp should be on January 2020.
- Removing data oscillations: oscillations, also known as data jumps, are wrongly reported locations. They may seriously distort the travel features by mistakenly showing that a device suddenly jumps to a faraway place. The identification and removal of data oscillations requires a reasonable algorithm differentiating between data jumps and normal points.
- Standardizing the datasets: the data are formatted with the following attribute names and data types - device_id (string), utc_timestamp, i.e., unix time (integer), latitude (float), longitude (float), and raw_speed (float).

All steps except removing data oscillations are straightforward. The step of removing data oscillations is separately discussed in the next section.

Chapter 4: Identification and Removal of Data Oscillations

Data oscillations present wrong or biased mobility patterns. For data driven research, identifying and removing data oscillations plays a critical role in the data preparation pipeline. As discussed in the literature review, previous studies show various limitations and deficiencies. Currently, the pattern-based methods have the greatest potential to be quantified for GPS data cleaning in this research. Although they are designed for cellular data, which present the ping-pong phenomenon caused by the change of cellular towers to which a mobile device connects, the core ideas of these methods are of significance. By these methods, data oscillations are defined as points showing abnormal movement patterns of location sequences and usually show impossibly high travel speeds (Wu et al., 2014; Lovan et al., 2013; Wang & Chen, 2018; Lee & Hou, 2006). The state-of-the-art method is proposed by Wu et al., 2014, who design several heuristics for identifying data oscillations from cellular data. In this study, their method is revised and expanded for the application to the nationwide raw truck GPS data from the following aspects:

1. A flexible spatial-temporal formulation is constructed by using adaptive instead of static parameters, which includes more cases of data oscillations;
2. Location uncertainty is reduced by using level-7 geohash zones instead of raw latitude and longitude coordinates;
3. Inspired by “stable period”, the definition of “stable community” and “stable zone” is proposed to label the points that are believed to be normal points,

which is more flexible by building local communities is more reliable by additionally considering dwell time and frequency;

4. More movement patterns of data oscillations are included by additionally considering data oscillations that continuously occur one-by-one;
5. Parameters are chosen through sensitivity tests or empirical experience from literature review instead of being arbitrarily decided.

Data oscillations present unreasonable movements. In order to quantify these abnormal movements or reduce location uncertainty, studies usually transform the raw latitude and longitude location into clusters (Wang and Chen, 2018). Considering the computation cost of spatial clustering, another way of aggregating the location data - geohash zone system - is applied, which is generated very fast. Instead of latitude and longitude coordinates, a geohash zone represents the position of a GPS sighting. This greatly reduces the uncertainty of the GPS points' locations and serves as the basic unit for formulating the movement patterns of data oscillations. Specifically, raw GPS sightings are projected to level-7 geohash zones (152.9 meters × 152.4 meters). Level-6 geohash zones (1.2 kilometers x 609.4 meters) are too large to capture data oscillations and level-8 geohash zones (38.2 meters x 19 meters) are unnecessarily granular. Hence, the raw GPS trace of each truck vehicle is denoted by a sequence of level-7 geohash zones. Abnormal movement patterns are identified from these simplified movements and data oscillations are removed accordingly.

4.1. Stable Communities and Stable Zones

A major assumption is proposed that if a vehicle is frequently observed (multiple GPS sightings) or is observed long enough (long dwell time) at a location, this location is believed to be a true visit. The true here does not refer to a level-7 geohash zone. Instead, a dynamic community is built to represent a location since vehicles are not just staying static at a place and in most cases are moving toward further destinations. The community here is a temporary group of continuous sightings for a given device during a time period. Several different communities can form for a device. Each community grows by including more and more continuous sightings that are close enough, i.e. less than $dist_c$ miles between two continuous sightings. Since the traces of a device are represented by a sequence of level7 geohash zones, a community built upon this is also a group of level-7 geohash zones. Sometimes, a community can only contain a single geohash zone including one or multiple GPS points. The value of $dist_c$ reflects the level of tolerance of data oscillations. For example, data oscillations that happen within the range of $dist_c$ are included as a part of a community and may not be identified as data oscillations in the end.

Note that a community could be a group of data oscillations. Based on the investigation of the raw sightings, when one data oscillation occurs, additional oscillations could occur around the first oscillation. For the purpose of differentiating true communities and oscillation communities, the definition of a stable community is proposed, which means that if a vehicle presents enough occurrences or dwells long enough in a community, this community is believed to be a true visit and is determined to be a stable community. All geohash zones included by a stable community are

defined as stable zones and all GPS points in a stable community are determined to be true visits. Each community has two attributes: frequency and duration. Frequency is the total number of sightings and duration is the time interval between the first and the last sighting. The threshold of frequency is determined to be 5 occurrences since the analysis of the raw sightings shows that the mean value of sightings in each level-7 geohash zone is 5. The duration threshold is determined to be 300 seconds since it is a state-of-the-practice dwell time value used for determining trip stops from GPS data. The sensitivity analysis in Figure 2 also shows that the number of identified oscillations peaks when the duration threshold value is set as 300 seconds.

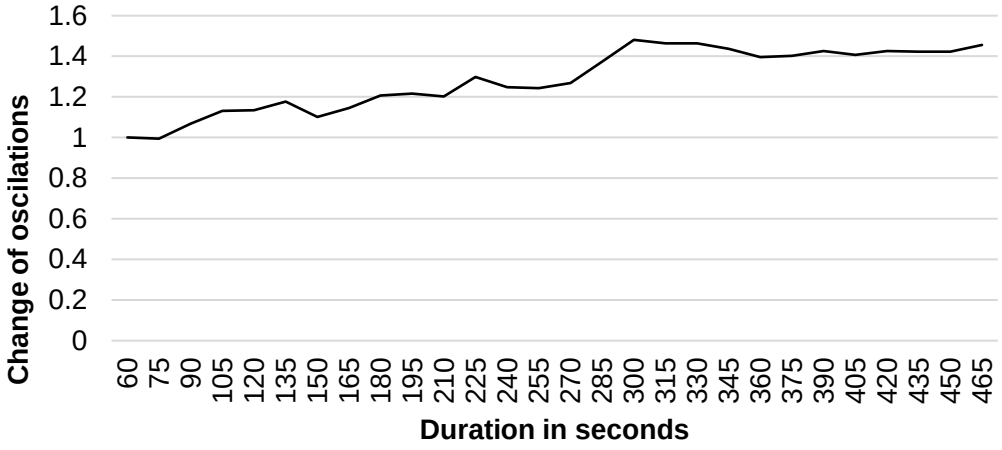


Figure 2. Sensitivity analysis of the duration threshold.

4.2. Bi-level Heuristics for Identifying Data Oscillations

Based on the definition of stable communities and stable zones, two heuristics at zone level and at community level separately are developed to identify data oscillations. The inspiration of the heuristics comes from Horn et al., 2014 that data oscillation moves to a relative faraway place at an abnormal speed. There are some revisions and improvements on the algorithms. First, Horn et al., 2014 uses clusters to

construct movement patterns while in this study communities and level-7 geohash zones are utilized as the representation of the locations, which is more computationally efficient when dealing with large-scale datasets. Second, Horn et al., 2014 uses static values of spatial-temporal parameters while in this study the corresponding heuristic is designed to use dynamic values. This makes the heuristic more flexible, so as to identify more cases of data oscillations. Third, the idea of stable communities and stable zones is developed and heuristics are accordingly revised to identify oscillations. Four heuristics are designed to identify data oscillations at geohash7 level (Heuristic 1a, 1b) and at community level (Heuristic 2a, 2b), which are described as below.

Heuristic 1a: It is assumed that a vehicle cannot finish a round trip within an extremely short time interval. For example, if a vehicle, within 30 seconds, starts from a stable level-7 geohash zone and returns to the same level-7 geohash zone that is also in stable status, all the middle sightings between these two geohash zones are determined as data oscillations, and removed. The threshold of 30 seconds is calculated as $t_{min} = 2 * dist_c/v_{max}$, where $dist_c$ = 0.5 miles, $v_{max} = 155/1.3 = 120\ mil/hr$. $dist_c$ is used since the middle geohash7 zone(s) are probably unstable zones, which have to be at least $dist_c$ far away. Otherwise, the device is in the same stable community as its previous community. 155 miles per hour is the twice of the allowed driving speed (Horn et al., 2014). Since the geodesic distance instead of road network distance is used here, a detour factor of 1.3 is used for making up for the difference.

Heuristic 1b: Considering that the GPS data are infrequent in some cases, it might be too demanding to satisfy the detection criteria mentioned above. Thus, a

supplementary check is conducted. It is assumed that a vehicle cannot move to a faraway place at an abnormally high speed for ground transportation. For any pair of two consecutive level-7 geohash zones, if one is stable while the other is unstable, the unstable zone is investigated to see if any oscillation exists. For example, if the distance between the two zones exceeds 5 miles and the time interval between the two zones is less than 2.5 minutes, then all sightings in the unstable zone are determined to be data oscillations and removed. Instead of a single speed parameter, distance and time interval are used, because drifting GPS points are common which may show very high speed but are actually not true data oscillations. These drifting points do not have bad effects on the inferences as data oscillations. Wu et al., 2014 exemplify this heuristic by using 5 km and 1 minute for cellular data, which is too high for GPS data. Here, the distance threshold of 5 miles and the time threshold of 2.5 minutes are arbitrarily determined. The corresponding speed is 120 miles per hour, which is the maximum allowed speed with the detour factor considered.

Heuristic 2a: the oscillations considered here are those that are geometrically obvious. Any sequence of three consecutive communities of a device can form a triangle. The distance and speed between communities are checked against two conditions as shown by equations (1) & (2).

$$v_{1-2} * v_{2-3} > 155 * 155 \quad (1)$$

$$dist_{1-3} < 0.25 * \min(dist_{1-2}, dist_{2-3}) \quad (2)$$

where v_{1-2} and $dist_{1-2}$ are the speed and the distance respectively from the first community to the second community; similarly, v_{2-3} and $dist_{2-3}$ are the speed and the distance from the second community to the third community; $dist_{1-3}$ is the distance from the first community and the third community. All GPS points in the middle community are probably oscillations and are removed.

Heuristic 2b: sometimes, such oscillations as described by heuristic 3 can continuously occur. When there are multiple continuous such data oscillations, true oscillations must be decided. Starting from the first community, all such communities satisfying the two conditions above are labeled with either an odd or an even position number. Either odd or even communities are removed. It is assumed that the communities with shorter dwell times are the true oscillations and are removed. If the odd and even communities have the same dwell time, either one is removed.

4.3. Results

The four heuristics define four types of data oscillations. The percentage of data oscillations by each type is summarized in Table 3. After data cleaning, 2.4% of data points removed.

Table 3. Percentage of identified data oscillations by types.

Data Oscillation Types	Percentage Out of All Points
Type 1a	0.00007%
Type 1b	0.00042%
Type 2a	2.38288%
Type 2b	0.01972%

Sum	2.40309%

The proposed heuristics can be applied to the data oscillation removal of not only the truck GPS data but also other types of location data, such as mobile location data collected by RFID, Wi-Fi, cellular tower, Bluetooth, GPS, and LBS (location based service). In the transportation field, another major application of the location data is in passenger travel behavior analysis.

Chapter 5: Vehicle Type Classification

The reasons for conducting vehicle type classification in this study is for the bias analysis of vehicle type composition and for the differentiated parameter specification of the trip identification algorithm by weight classes. First, GPS data are a sample from the real world. Various biases can exist. Vehicle type bias is one of the major biases since the GPS dataset is usually multi-sourced from various truck companies. Dataset I GPS data is collected from private cars and commercial vehicles. Dataset II GPS data mainly contain heavy-duty trucking fleets. Both datasets I and II are utilized in this study for more complete truck mobility information. The vehicle type analysis in this study not only serves as the basics of deriving national truck travel demand at population level but also provides insights into other truck-related studies by truck types, such as truck travel distance analysis and truck parking analysis. Ignoring the unbalanced feature of the dataset would lead to wrong and biased travel statistics. In this study, a weighting process specifically for vehicle types is conducted later to produce representative results. Second, vehicle type classification is done for a better deployment of trip identification algorithm. The trip identification algorithm utilized in this study is a trip-end-based algorithm relying on a set of spatial-temporal parameters. Considering that vehicles of different weight classes present different spatial-temporal behavioral features, the input parameters, i.e., a set of time, distance, and speed thresholds, of the algorithm should be investigated for better performance. For example, light trucks mainly provide delivery or distribution services in nearby areas. They may have more quick stops due to the traffic signals or frequent short-term congestions, lower cruising speeds, and a shorter headway between vehicles than long-

haul heavy trucks. The loading and unloading time is different between some heavy trucks and light trucks. Hence, it is important to discuss the parameter values of the trip identification algorithm for different truck types.

5.1. Methodology

As detailed in the literature review, a Random Forest (RF) algorithm is selected for classifying vehicles. It is a powerful classifier with high accuracy in various classification problems. It yields the labeling results based on an ensemble of decision trees. The theory behind this is ensemble learning, which utilizes the results of multiple models. It proves to produce more accurate predictions than any single model (Fawagreh et al., 2014). Bootstrap Aggregation / Aggregating, also called as Bagging, is an ensemble machine learning algorithm. Random forest is a type of Bootstrap Aggregation algorithm. Like the general idea of Bootstrap Aggregation, it first randomly selects sample sets with replacement (step 1); then each sample set is used for modeling by a decision tree (step 2); lastly, the predictions of all decision trees are aggregated by a voting process and the prediction with the highest voting score is selected as the final prediction (step 3). It is different from the general Bootstrap Aggregation methods since it additionally applies feature bagging when modeling decision trees. Only a limited number of features can be used for each decision tree modeling, to avoid building trees that are similar to each other. The feature bagging process and the bootstrap sampling reduces the dependence between the decision trees. Random forest is much more robust and stable than a decision tree. With random forest, the variance of decision tree models is reduced by randomly selecting the input data of each decision tree in a bootstrapping way and by averaging the predictions of all

independent models. With so many different and highly uncorrelated decision trees built in a random forest, the overfitting which is concerning in a single decision tree is usually not a problem in a random forest. Additionally, random forest runs very fast since all decision trees are modelled in parallel. The details of each step are described as follows.

Step 1: constructing sub-datasets by bootstrapping. Bootstrap is a common statistical method for making an estimate of a population by averaging the estimates from multiple small data samples in such a way that these small data samples are constructed by random sampling with replacement. With replacement, each selection of samples is fully random. Hence, each decision tree modeling is fully independent from others and the prediction accuracy is increased.

Step 2: modeling by decision trees. A decision tree consists of decision nodes (white solid circles), branches (arrows connected two nodes), and leaf nodes (black solid circles). The first decision node (e.g. X1 < alpha) is also called as root node. At each decision node, one feature is selected and checked against a condition. The feature can be continuous or binary. At each decision node, there can be two or more branches. The same feature can be repeatedly selected by different decision nodes. For a classification problem using a decision tree, the feature selected at each decision node is decided by the information gain (IG). It measures the amount of information that a feature provides about a category. It is computed by the change of entropy supposing the dataset is partitioned by a feature. By splitting, the dataset is divided into two or more sub-datasets. For each sub-dataset, an entropy value can be calculated. Suppose

that the dataset contains n categories in total and, for any sub-dataset j from l total sub-datasets, the probability of selecting a sample from category i from n possible categories is p_i. Then the entropy, as a randomness measurement that quantifies the impurity of a sub-dataset, is calculated by $E_{j=1,2,..,m} = -\sum_{i=1}^{n} p_i log_2^{p_i}$. The impurity of a decision node is measured by the weighted entropy from all sub-datasets: $E_w = \sum_{j=1}^{l} w_j E_j$. The number of samples of each sub-dataset determines the weights: $w_j = n_j/(n_1 + n_2 + \cdots + n_j + \cdots + n_l)$. Entropy ranges from 0 to 1. If the dataset is balanced with equal numbers of samples from different categories, the entropy is 1, meaning the highest impurity; if the dataset only has one category, the entropy is 0, meaning complete purity. The initial entropy of a tree is 1. With splitting by a feature, the entropy change (i.e. entropy loss) is defined as the information gain. A high information gain is desired so that the tree's entropy is highly reduced. Finally, an optimal status of the tree's splitting is reached when the weighted entropy of the tree is very close to zero. At the status, each sub-dataset is very pure, with most of the samples belonging to a single category. The tree is always branched by a feature that has the most entropy change / loss (i.e. the largest information gain) at the moment. Thus, the order of features selected by the decision nodes shows their relative importance with regards to information gain. In other words, the entire input dataset for a given decision tree is subdivided by multiple sets of branches with each set starting from the root node to a leaf node. The depth of a decision tree is the number of branches along with the longest path (i.e. the largest set of branches) and the size of a decision tree is the number of nodes in the tree. This process of subdividing the input dataset and growing a set of branches is known as binary recursive partitioning. It continuously expands the tree

until the data within each subdivided data space is homogenous enough (e.g. all data points in the current data space have the same category) or some other criteria (e.g. the depth of the tree is limited) is satisfied. Then a leaf node is reached. A leaf node denotes a classification result decided by a data space refined by a set of branches. Given an unlabeled data input, the data space that it falls into determines its classification label. In a random forest, there are a number of decision trees that are independently built in parallel. Each decision tree is built with a limited number of features, which is called feature bagging. Two major parameters in this step are the number of features for each decision tree and the number of decision trees. A rule-of-thumb is that the number of features that can be used for splitting in a decision tree is decided by $m = \sqrt{p}$, where p is the number of available input features. Random forest works better if the correlation between the decision trees is low. For this purpose, the feature that performs well will be internationally avoided for other trees. The second parameter can be decided based on test results.

Step 3: aggregating the predictions of all decision tree models by a voting process. A commonly used voting method for a classification problem is the majority method.

5.2. Data Segmentation

Although on average there is a clear difference between the two datasets, they have trucks that share similar GPS data qualities. This occurs because both datasets are passively-collected and mixed GPS data collection techniques are used. In other words, even within the same dataset (either I or II), trucks differ from each other regarding the GPS data quality. For a data-driven study, it is important to consider the disparities

within the raw data. Data segmentation through the K-means clustering method is therefore deployed. Input features are listed in the Table 4. In reality, each truck is in either static or moving status. Passive truck GPS data may not capture both states. To be usable for deriving trips, a truck should provide observations at least in moving status. Trucks that only provide static observations are removed. The trucks that provide static points in addition to moving points are also useful since in such case static points become meaningful for indicating the stop status. Therefore, two cases are jointly considered: (1) all observations, whether static or moving, are considered; (2) only moving observations are considered.

Table 4. Data features used for data segmentation.

Data features	Statistics of logging frequency	Mean, 25, 50, 75, and 95 percentile of the frequency: $f_{avg,j}$, $f_{p25,j}$, $f_{p50,j}$, $f_{p75,j}$, $f_{p95,j}$. Frequency is the time interval between two consecutive points: $f_i = \Delta t_{i-1,i}$. j denotes the two cases above.
	Daily frequency	Average daily number of observations for each device: $\overline{obs}_{dy,j}$.
	Hourly frequency	Average hourly (date and hour) number of observations for each device: : $\overline{obs}_{dyhr,j}$

For determining the optimal way of clustering, eight experiments (k=2, 3, …, 9) are conducted. A Silhouette Score (Rousseeuw, 1987) is calculated to measure the clustering performance, i.e., how the data points in a cluster are similar to each other and how they are different from points in another cluster. The changes of Silhouette Score are plotted in Figure 3, along with which two more groups corresponding to only case (1) and only case (2), respectively, are added for comparison. Figure 3 shows that jointly considering (1) and (2) having four clusters yields a much-higher-than-others score, as shown by the line and the dark bar plots. This optimal score is very close to 1

indicating an excellent clustering result. In comparison, if only considering (1) or (2), having two clusters is sufficient.

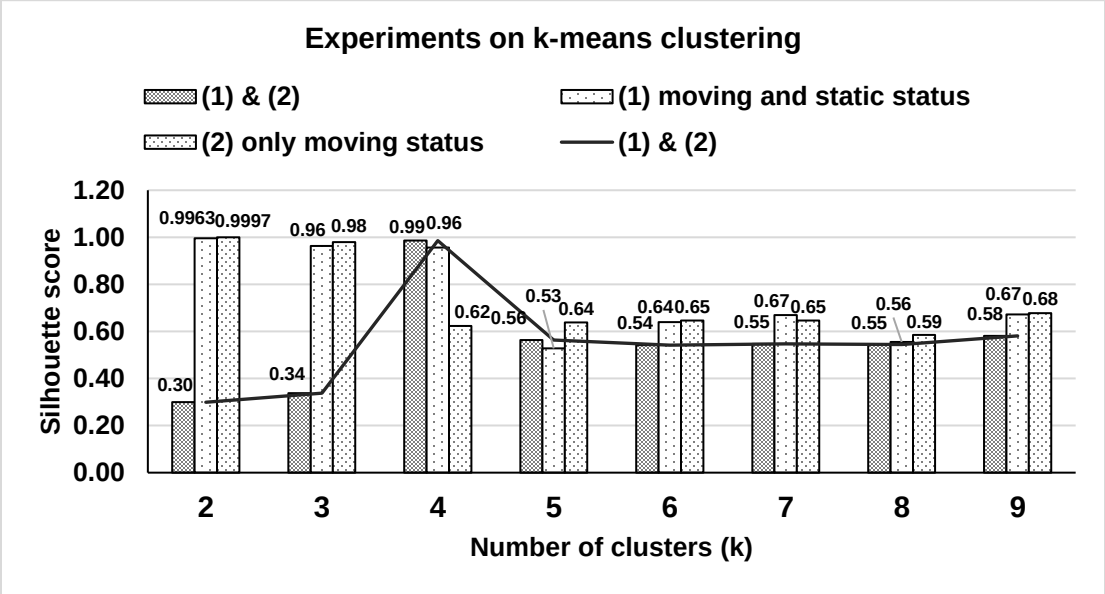


Figure 3. Experiments on K-means clustering.

After clustering by $k = 4$, the distribution by cluster, data source, and truck type, is summarized in Table 5. It shows that three clusters (0, 1, 2) contain trucks from both datasets and cluster 3 only has trucks from dataset II. All trucks within the same cluster should have similar data features. Therefore, in each cluster, dataset I trucks are used for training a RF model as described in the next section. Then the trained model is applied to dataset II for truck classification. For cluster 2 and 3, dataset I does not provide sufficient training data, so the RF classification is not applied to the corresponding dataset II. This is a negligible data loss (0.0031%).

Table 5. Distribution of trucks by cluster, data source, and truck type.

Cluster	Data source	Truck type	Proportion	Sum
0	Dataset I	Heavy	0.4377%	99.7306%
	Dataset I	Light-medium	83.5331%	

	Dataset II	N/A	15.7597%	
1	Dataset I	Heavy	0.0063%	0.2637%
	Dataset I	Light-medium	0.1545%	
	Dataset II	N/A	0.1029%	
2	Dataset I	Heavy	0.0001%	0.0056%
	Dataset I	Light-medium	0.0025%	
	Dataset II	N/A	0.0030%	
3	Dataset II	N/A	0.0001%	0.0001%

5.3. Implementation and Results

Four groups of features are explored in this study and summarized in Table 6. The first group includes cruising features that prove to be significant in current studies, including speed, acceleration, and heading direction change. Heading direction is usually actively collected while passive GPS data do not include this information. Instead of heading direction, angular change and speed are explored. The second group includes newly designed activity features including mobility radius and nighttime factor. Light trucks mainly serve areas within 50 miles while heavy trucks may serve area 200 miles away. Moreover, as discussed by Ma et al., 2012, light trucks are more active during daytime while heavy trucks are more active during nighttime. The distribution plot of average freight weight per truck by hour of the day in their research shows that 7am-5pm is the period with the lowest average weight per truck. Hence, a nighttime factor is designed. The third group includes built-in environment factors. It is assumed that there is a difference in the land use pattern between truck types. For instance, local trucks are mainly active in urban areas while heavy trucks may spend much more time on less-populated area. Built-in environment features – street intersection density and transit frequency are therefore incorporated. Correlation

analysis in Appendix A shows that many features in the cruising group are highly correlated. For lower feature redundancy, lower dimensionality, and higher interpretability, features in cruising group are sub-grouped and for each group Principal Component Analysis (PCA) (Jolliffe, 2002) is implemented to produce two principal components. Not every feature is necessarily useful or informative. PCA reduces data noise while keeping the maximum information (feature variance). For each subgroup, the principal components with at least 10% explained variance are selected as the input feature of RF model. For comparison, the union of such principal components from the two clusters is finally used. In total, 15 principal components are used, which are denoted as $pc_j\_g_k$ (the 1st, 2nd, …, jth principal component from group g_k). The PCA report of selected principal components is in the Appendix B, which contains the percentage of explained variance and loading matrix by each principal components.

Table 6. A summary of input features.

Group	Input feature	Description	Sub-group
Cruising	Statistics of interval speed	Mean, 25, 50, 75, and 95 percentile of the interval speed: $\tilde{v}_{avg}$, $\tilde{v}_{p25}$, $\tilde{v}_{p50}$, $\tilde{v}_{p75}$, $\tilde{v}_{p95}$. Interval speed is the average speed between the current point and its previous point: $\tilde{v}_i = \frac{\Delta d_{i-1,i}}{\Delta t_{i-1,i}}$	g_1: features related to speed
	Statistics of point speed	Mean, 25, 50, 75, and 95 percentile of the point speed: $\dot{v}_{avg}$, $\dot{v}_{p25}$, $\dot{v}_{p50}$, $\dot{v}_{p75}$, $\dot{v}_{p95}$. Point speed is an attribute of the raw data.	
	Statistics of acceleration based on interval speed	Mean, 25, 50, 75, and 95 percentile of the acceleration based on interval speed: $\tilde{a}_{avg}$, $\tilde{a}_{p25}$, $\tilde{a}_{p50}$, $\tilde{a}_{p75}$, $\tilde{a}_{p95}$. Acceleration is defined as the absolute value of the difference in interval speed divided by the time interval: $\tilde{a}_i = \frac{\tilde{v}_i - \tilde{v}_{i-1}}{\Delta t_{i-1,i}}$	g_2: features related to acceleration
	Statistics of acceleration based on point speed	Mean, 25, 50, 75, and 95 percentile of the acceleration based on point speed: $\dot{a}_{avg}$, $\dot{a}_{p25}$, $\dot{a}_{p50}$, $\dot{a}_{p75}$, $\dot{a}_{p95}$. Acceleration is defined as the absolute value of the difference in interval speed divided by the time interval: $\dot{a}_i = \frac{\dot{v}_i - \dot{v}_{i-1}}{\Delta t_{i-1,i}}$	

	Statistics of the difference in interval-speed-based acceleration	Mean, 25, 50, 75, and 95 percentile of the difference in acceleration based on interval speed: $\tilde{a}'_{avg}$, $\tilde{a}'_{p25}$, $\tilde{a}'_{p50}$, $\tilde{a}'_{p75}$, $\tilde{a}'_{p95}$. Acceleration is defined as the absolute value of the difference in interval speed divided by the time interval: $\tilde{a}'_i = \frac{\tilde{a}_i - \tilde{a}_{i-1}}{\Delta t_{i-1,i}}$	g_3: features related to the difference of acceleration
	Statistics of the difference of point-speed-based acceleration	Mean, 25, 50, 75, and 95 percentile of the difference in acceleration based on point speed: $\dot{a}'_{avg}$, $\dot{a}'_{p25}$, $\dot{a}'_{p50}$, $\dot{a}'_{p75}$, $\dot{a}_{p95}$. Acceleration is defined as the absolute value of the difference in point speed divided by the time interval: $\dot{a}'_i = \frac{\dot{a}_i - \dot{a}_{i-1}}{\Delta t_{i-1,i}}$	
	Statistics of angular change	Mean, 25, 50, 75, and 95 percentile of the absolute angular change: $\Delta\theta_{avg}$, $\Delta\theta_{p25}$, $\Delta\theta_{p50}$, $\Delta\theta_{p75}$, $\Delta\theta_{p95}$. Angular change is measured by the absolute angle between two lines – one is from the current point to the next point, the other is from previous point to the current point: $\Delta\theta_{i-1,i}$	g_4: features related to the angular change
	Statistics of angular speed	Mean, 25, 50, 75, and 95 percentile of the angular speed: ω_{avg}, ω_{p25}, ω_{p50}, ω_{p75}, ω_{p95}. Angular speed is defined as: $\omega_i = \frac{\Delta\theta_{i-1,i}}{\Delta t_{i-1,i}}$	g_5: features related to angular speed
Activity	Mobility radius	The maximum distance from the center point: r	N/A
	Nighttime factor	Percentage of GPS records being observed during nighttime (the time except 7am-5pm) in moving status (non-zero point speed): $ntf = \sum_n^N w_n D3B_n$.	
Built-in environment	Street intersection density	Weighted mean of D3B (Street intersection density weighted, auto-oriented intersections eliminated) by census block group (CBG): $d3b = \sum_n^N w_n D3B_n$.	
	Transit frequency	Average D4E (Aggregate frequency of transit service per capita) by CBG: $d4e = \sum_n^N w_n D4E_n$.	

For each cluster, balanced data from dataset I with 50% light-medium trucks and 50% heavy trucks are used as RF input with a training-testing split ratio of 7:3. A five-fold cross validation method is applied to avoid overfitting and ensure the generalization ability of the trained model. Three hyperparameters of the RF model are considered: ntree (number of trees), max_depth (maximum depth of the tree), and max_features (number of features considered at each leaf). Based on a rule of thumb, max_features is set as the square root of number of input features. Then, ntree and max_depth are determined by experiments. In the model, the heavy type is labeled as positive. Considering the relatively limited volume of heavy trucks and the fact that

dataset II is dominated by heavy trucks, a Type II error (false negative) should be as rare as possible, and so a high recall score is preferred. Besides, sensitivity analysis is conducted to show how different hyperparameters influence the prediction result. For each experiment, the input is a set of hyperparameter values for cluster 0 and 1, and there are two outputs: (1) the recall score based on testing data; (2) the percentage of predicted light-medium trucks in dataset II. The results of experiments are presented in Figure 4, in which a two-period moving average trending line of (2) is added. The black dotted trending line clearly shows that experiment 12 is the "elbow" point with 5.50% predicted light-medium trucks. And as the green line shows, experiment 12 has the second highest recall score, i.e. 94.44%, which is just slightly lower than the highest recall score of 95.00%. A ratio of 5.5% is a reasonable value since it is similar to an estimate around 5% of non-heavy trucks in dataset II in another study (Zanjani et al., 2015) using a different method. Therefore, the set of hyperparameters in experiment 12 is finally adopted.

The hyperparameter value and confusion matrix is summarized in Table 7. The confusion matrix demonstrates model performance on testing data. With exactly the same input features, the RF model of cluster 0 clearly works better. As previously discussed, cluster 0 corresponds to high-frequency data while cluster 1 corresponds to low-frequency data. This might be the reason for the different RF model performances.

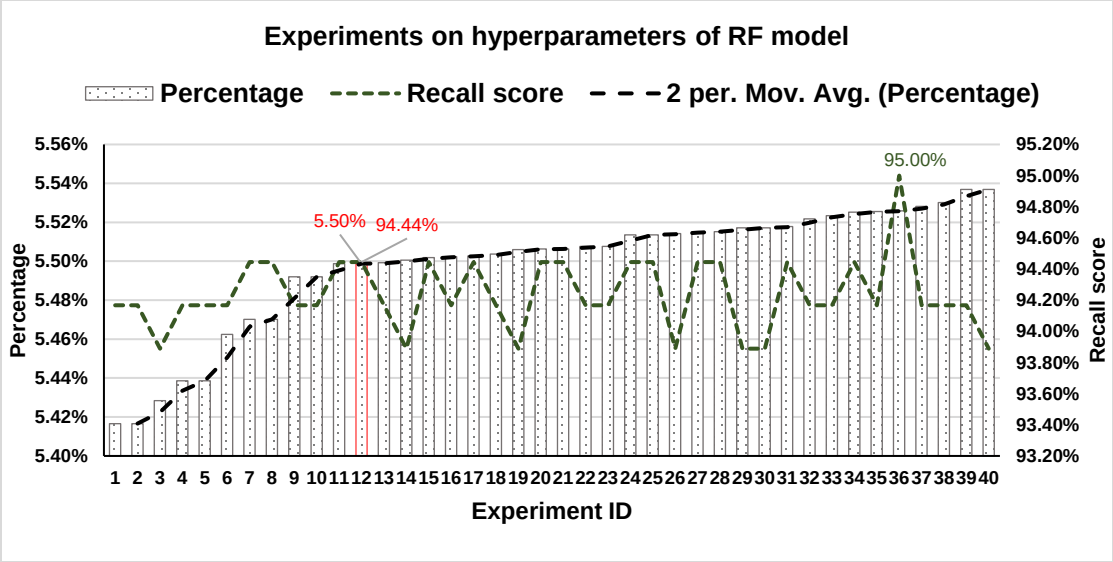


Figure 4. Experiments on hyperparameters of RF model.

Table 7. RF model hyperparameter and confusion matrix.

Cluster	ntree	max_ depth	max_ features	Confusion matrix		Predictions	
						Light-medium	Heavy
0	40	10	4	True	Light-medium	96.5%(303)	3.5%(11)
					Heavy	2.1%(6)	97.9%(280)
1	60	20	4		Light-medium	87.3%(69)	12.7%(10)
					Heavy	18.9%(14)	81.1%(60)

Furthermore, the SVM model is set as the baseline model. It is compared with the RF model from four aspects: Receiver Operating Characteristic (ROC) curve and Area under the ROC curve (AUC), recall score, accuracy, and precision as shown in Figure 5 and Table 8. All these measures are computed from the testing data. First, the five-fold cross validation in Figures 5(a)-(d) shows that the ROC curve and AUC value of each fold are very similar to each other, which means that there is no overfitting issue and the trained model has a high generalization. Second, Figures 5(a) & (c) show that RF has very good classification performance with a high AUC value. Third, RF is not inferior to SVM and actually performs slightly better than SVM. Regarding recall

score, accuracy, and precision, RF yields similar result as SVM for cluster 0 (high-frequency data) while it is obviously better than SVM for cluster 1(low-frequency data). In the literature, the SVM developed by Sun and Ban, 2013 has 0.958 accuracy in classifying vehicles from GPS data. The deep learning model of Simoncini et al., 2018 has an AUC value of 0.939 for low-frequency GPS data. With SVM as baseline model, this study demonstrates the high capability of RF model for conducting GPS-based vehicle classification in both high- and low- frequency scenarios.

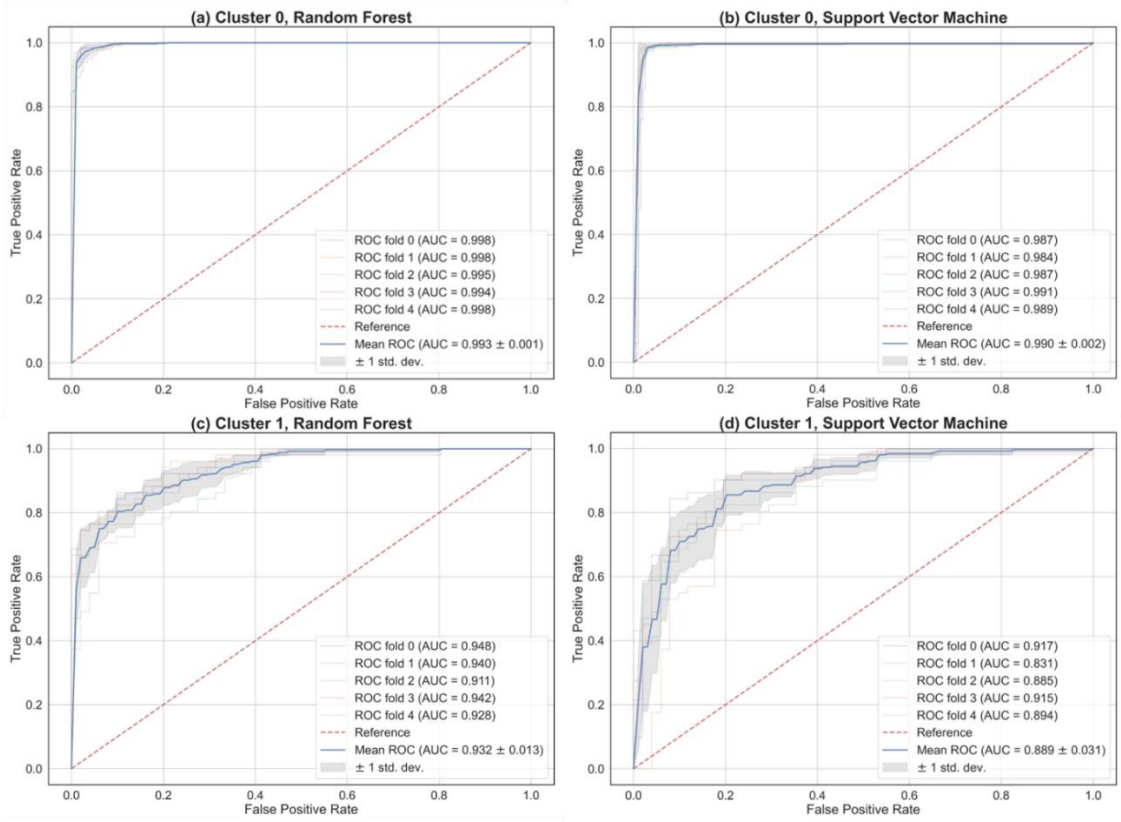


Figure 5. Five-fold ROC curves for RF and SVM.

Table 8. Accuracy measures of RF and SVM.

Cluster	RF			SVM		
	Recall	Accuracy	Precision	Recall	Accuracy	Precision
0	0.9790	0.9717	0.9622	0.9790	0.9783	0.9756
1	0.8108	0.8431	0.8571	0.7568	0.7712	0.7671

Although the same set of input features is used for both clusters, the RF performance varies greatly. It is important to understand how each input feature contributes to the prediction in these two different scenarios. Figure 6 shows the importance measure of input features. For cluster 0, i.e., high-frequency GPS data, acceleration-related features (g_2) and difference-in-acceleration-related features (g_3) contribute most. $pc_{1\_}g_2$ ranks 1st and $pc_{1\_}g_3$ ranks 2nd.. Sun and Ban, 2013 also finds that acceleration-related features are very predictive. According to the loading matrix, $pc_{1\_}g_2$ mainly accounts for the variance of the mean, 75th, and 95th percentile of point-speed-based acceleration ($\dot{a}_{avg}$, $\dot{a}_{p75}$, $\dot{a}_{p95}$) and $pc_{1\_}g_3$ mainly accounts for the mean of difference in point-speed-based acceleration ($\dot{a}'_{p50}$, $\dot{a}'_{p75}$, $\dot{a}'_{avg}$). The 3rd important feature - $pc_{1\_}g_5$ – almost equally accounts for the variance of ω_{avg}, ω_{p25}, ω_{p50} , ω_{p75} , ω_{p95} regarding angular speed. In the literature, the change of instantaneous heading direction is set as a predictor, which is actively collected. This study demonstrates that the speed of heading direction change based on two consecutive GPS points is also predictive. For passively collected GPS data, this is a good substitution for instantaneous heading change speed. In a high-frequency scenario, these mentioned features are highly recommended for RF model inputs. It is noted that features derived from interval speed do not show much importance so far and the four newly proposed features based on activity space and built-in environment are also relatively weak.

However, for cluster 1, i.e., low-frequency scenario, the three new features - transit frequency ($d4e$), mobility radius (r), and nighttime factor (ntf), play a

significant role. Since public transit usage and urban area size are highly correlated (Taylor et al., 2009), it is not surprising that transit frequency, as an indicator of urbanization degree of a truck's on-duty traces, ranks 1st. Yet, its importance is very similar to the second feature - $pc_1\_g_2$. Mobility radius and nighttime factor also exhibit their power of discriminating light-medium and heavy trucks with a rank of 4th and 7th respectively. Compared to these features derived from interval speed, the newly proposed features are less sensitive to data frequency and therefore more stable and reliable. Since a majority of cruising features is computed from two consecutive GPS points, high- and low- frequency GPS data present different patterns of such cruising features. Accordingly, the PCA result differs between cluster 0 and 1. Taking $pc_1\_g_2$ as an instance, in addition to three aforementioned features ($\dot{a}_{avg}$, $\dot{a}_{p75}$, and $\dot{a}_{p95}$), the mean and 95th percentile of interval-speed-based acceleration ($\tilde{a}_{avg}$ and $\tilde{a}_{p95}$) become important in cluster 1 as shown by loading matrix. In the low-frequency scenario, $pc_2\_g_3$ ranks 3rd, which mainly accounts for four features regarding the difference in interval-speed-based acceleration ($\tilde{a}'_{p25}$, $\tilde{a}'_{p50}$, $\tilde{a}'_{p75}$, and $\tilde{a}'_{p95}$) and three features regarding the difference in point-speed-based acceleration ($\dot{a}'_{p75}$, $\dot{a}'_{p95}$, and $\dot{a}'_{avg}$). In the high-frequency scenario, features derived from interval speed do not show much importance but some of them become highly important in the low-frequency scenario. The lower the frequency, the more difference there is between interval speed and point speed. As Simoncini et al., 2016 described, point speed is discriminative and sometimes noisy while interval speed is smooth and can be more reliable especially for low-frequency data. For the 5th and 6th important features of both clusters ($pc_2\_g_2$ and

$pc_{4\_}g_2$), interval-speed-based acceleration features also become significant. All important features mentioned so far are highlighted in the PCA report.

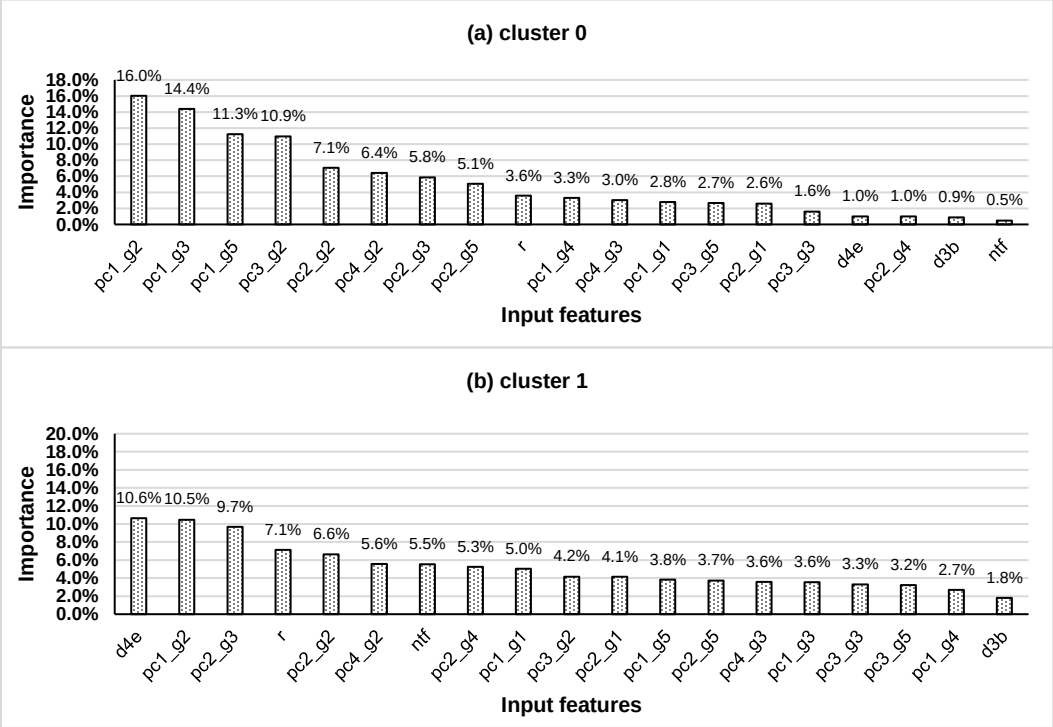


Figure 6. Feature importance measure.

For the three new important features, the marginal effect on the probability of being predicted as a heavy truck is demonstrated in Figure 7. Figure 7(a) shows that transit frequency positively influences the probability of being a heavy truck. Transit frequency denotes the urbanization degree of a truck's trajectories. It conforms to the fact that light-medium trucks have more trajectories in urban areas while heavy truck have more trajectories in less urban areas, such as inter-state highways. Figure 7 (b) shows that at first there is an increasing trend of being predicted as heavy with the increase of mobility radius and later on the probability flattens and is not much sensitive to the mobility radius. This makes sense since the threshold between light-medium and heavy trucks happens in the earlier stage. This threshold seems to be small. There are

three reasons. First, it is the radius of activity space. Second, Euclidean distance, instead of cumulative travel distance on roadway, is used. Trucks with multiple stops can have a very high detour factor. Third, passive GPS data have missing observations, so the most faraway stop may not be captured. Figure 7 (c) shows that the relative active strength during nighttime has a positive influence on being classified as a heavy truck. Overall, Figure 7 shows the sensitivity of these three new features. These features are less sensitive to data frequency, and thus can be applied to various GPS datasets.

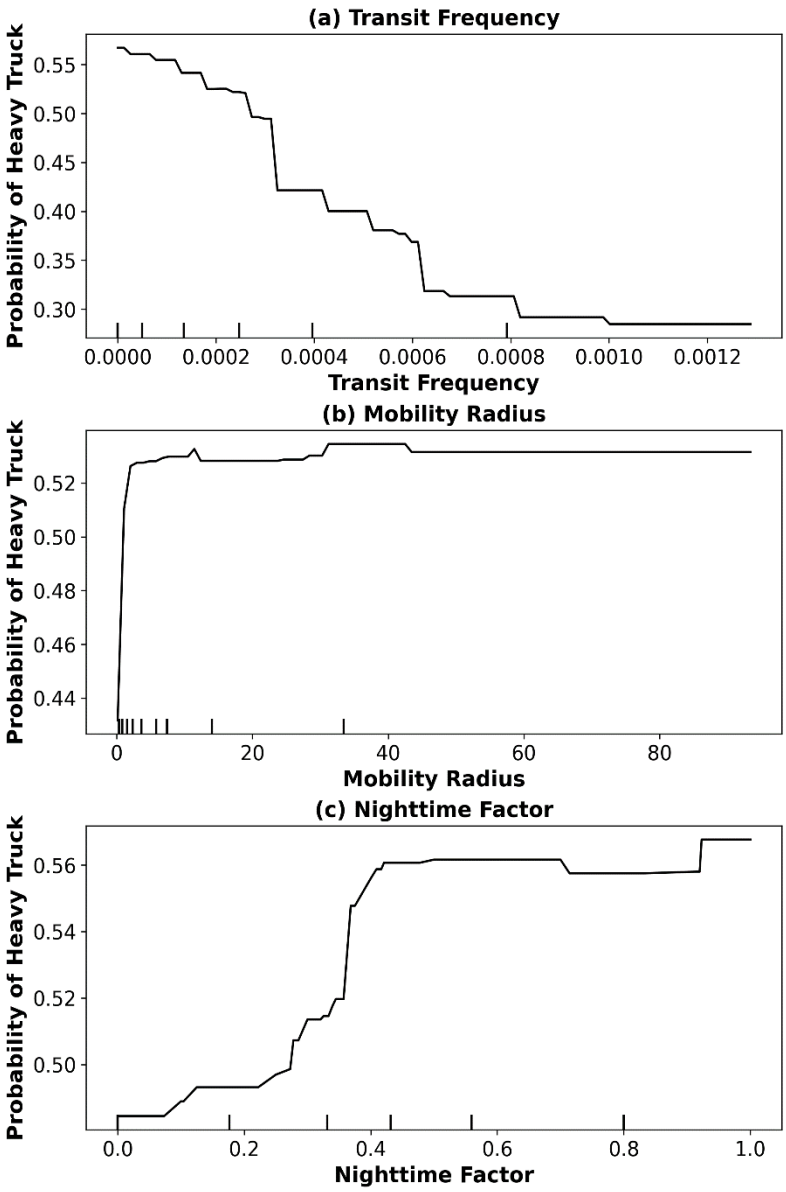


Figure 7. Partial dependence plots.

Chapter 6: Truck Trip Profile Preparation

6.1. Methodology

As discussed in literature review, state-of-the-practice methods for identifying trips from passively collected GPS data are threshold-based methods. Spatial-temporal parameters such as speed, dwell time, and distance are usually considered when designing the algorithm. In comparison to other types of methods such as clustering-based methods and machine learning methods, the threshold-based methods are more suitable for the application in this study for the following reasons. First, trucks of different weight classes show obviously different mobility patterns. An algorithm such as a threshold-based one has the potential of treating such differences by setting different threshold values for the spatial-temporal parameters. However, clustering-based methods and machine learning methods do not have such parameters that can specifically differentiate the trucks of different weight classes. Although both threshold-based and clustering-based methods necessitate predefined parameters, threshold-based methods have empirical references from the mobility patterns of truck types while clustering-based methods do not have such references. By applying a threshold-based method, the mobility differences are elaborated in this study to provide guidance for setting different threshold values for different truck types, which is missing in current applications. Second, machine-learning methods need validation dataset, which is unavailable for passively collected truck GPS data. Third, threshold-based methods identify the trips to the largest extent without data loss while clustering-based methods may identify some geographically independent points as outliers. For a

low-frequency GPS dataset, such geographically independent points (relatively more geographically isolated from other points) probably constitute a significant fraction, which should not be omitted.

There have been numerous threshold-based methods for identifying trips from GPS data, which have been performing well and have been widely applied in various studies. The major effort in this section is not to propose a different method. Instead, a threshold-based method is applied and its performance is investigated in the two scenarios: light-medium truck and heavy truck. In the literature, most applications of threshold-based algorithm arbitrarily decide the threshold values and no study has discussed how different threshold values will influence the model result. This study not only conducts many experiments on how identification results change with different sets of threshold values, but also compares the two scenarios by truck type. This is the major contribution of this chapter to the current research.

A recursive threshold-based algorithm for identifying trips from mobile location data is applied. The algorithm is developed from a previous work (Zhang et al., 2021) co-authored by the author, in which the algorithm is applied to passive GPS data to identify personal trips in multiple travel modes. In the present study, this algorithm is applied to identify truck trips from passive GPS data. This algorithm jointly considers the spatial-temporal relation between the current point, its previous point and its next point based on speed, distance, and dwell time. The algorithm's results fit well the passively collected location data by identifying trips in a conservative way with consideration of the missing points and low frequency. The

algorithm also has high applicability to passive truck GPS data. First, passive truck GPS data share very similar characteristics with passive passenger GPS data. Second, truck GPS data are less complicated and less noisy than passenger GPS data, since only a driving mode is involved. This study extends the application of this previously developed algorithm to the identification of truck trips. Moreover, parameter exploration by truck type is conducted. Light-medium and heavy trucks show obviously different spatial-temporal features. Such a difference and the sensitivity of parameters to this difference is explored in this study. Since dataset I has trip profiles, this algorithm is only applied to dataset II.

A brief description of the algorithm is as follows. The algorithm is applied at the truck level. Given a truck, all of its GPS points are ordered in time sequence with a default trip ID - 0 (meaning static status not belonging to any trip). Then the algorithm locates the first trip start point and checks all the following points. Each point remains in the default static status unless it is identified as (a) starting a new trip or (b) belonging to the same trip as its previous point. For each GPS point, the speed, distance, and time interval from its previous point ("speed from", "distance from", and "time from") or to its next point ("speed to") are used for deciding the status. Detailed conditions for updating the default static status to (a) or (b) are summarized below and plotted in Figure 8. The trip identification algorithm may identify some short truck trips that occur within a large establishment. Such short trips (e.g. less than 300 meters) are not useful and, hence, are removed.

(a) Starting a new trip: under one of the three conditions: (1) this current point is the first point. (2) its previous point retains the default static status. (3) its previous point is on a trip but this current point does not belong to the same trip as its previous point, if "speed to" >= speed threshold (s), this current point is labeled as the first point on a trip and is updated with a new trip ID;

(b) Belonging to the same trip as its previous point: under the condition that its previous point is already on a trip, there are two cases: (1) if "speed from" >= s, this current point belongs to the same trip as its previous point; (2) if "speed from" < s, "distance from" < distance threshold (d), and cumulative "time from" < time threshold (t), this current point is also identified to be on the same trip as its previous point. For both cases, this current point is updated with the same trip ID as its previous point. It should be noted that for case 2, if "speed from" < s is satisfied while others are not meet, this current point is further checked to see if it starts a new trip, i.e. status (a).

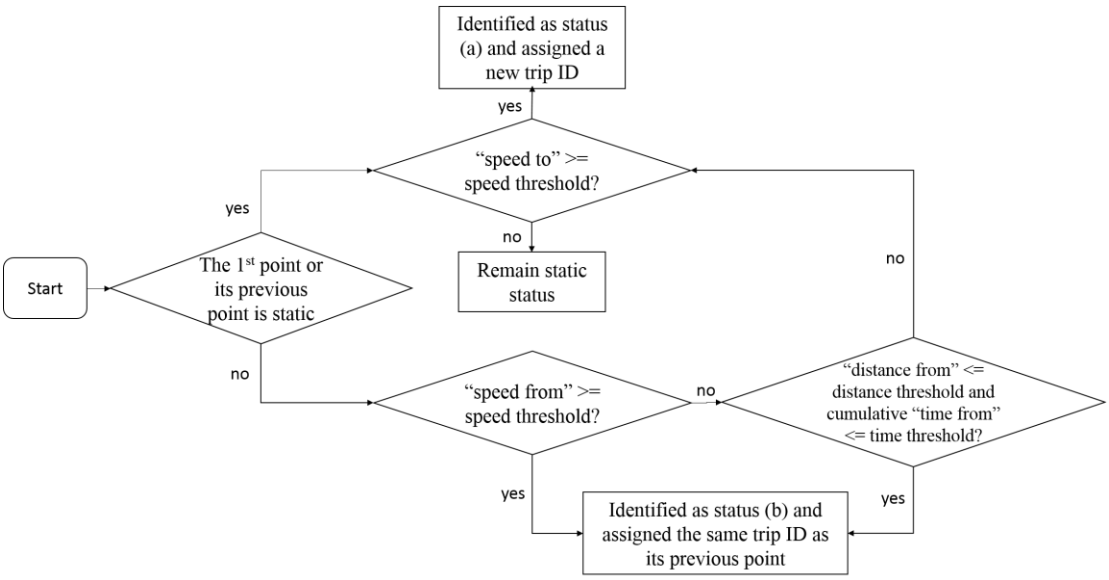


Figure 8. Flow chart of trip identification algorithm.

6.2. Specification of the Parameters

Although current studies use different threshold values when identifying trips from GPS data, they are usually arbitrarily decided or referenced from existing practices. In this study, comprehensive parameter tests are conducted to present the sensitivity of threshold values in different scenarios, i.e., light-medium versus heavy weight trucks, as presented in Figure 9. The 2020 FAF OD freight tonnage data by truck at the state level are used for Pearson correlation with the sample OD truck traffic aggregated from each parameter test result. On the one hand, a high Pearson's r is expected considering the similarity between truck travel flow and truck freight flow. On the other hand, FAF data have some deficiencies. First, 2020 FAF dataset is not ground truth; instead it is a forecasted product based on 2017 CFS and some other ancillary data. Second, FAF truck data mainly account for intercounty commodity flow. These intracounty flows that are usually under 50 miles are a data gap (FHWA, 2004,

Issues for Improving the FAF). Third, for light-medium- versus heavy- weight comparison in Figure 9, corresponding FAF data based on average weighted distance of shipment (< 100 miles and >= 100 miles respectively) are used. Such correspondence between truck weight class and shipment distance may introduce some error. Nevertheless, FAF is the best available data source for nationwide OD truck flow validation or calibration. Since the clustering process may bring about unmeasurable bias to the OD distribution, cluster-wise correlation with FAF is unreasonable; therefore, the threshold value sensitivity test is only conducted by truck type. Table 9 summarizes the general trends from test results in Figure 9.

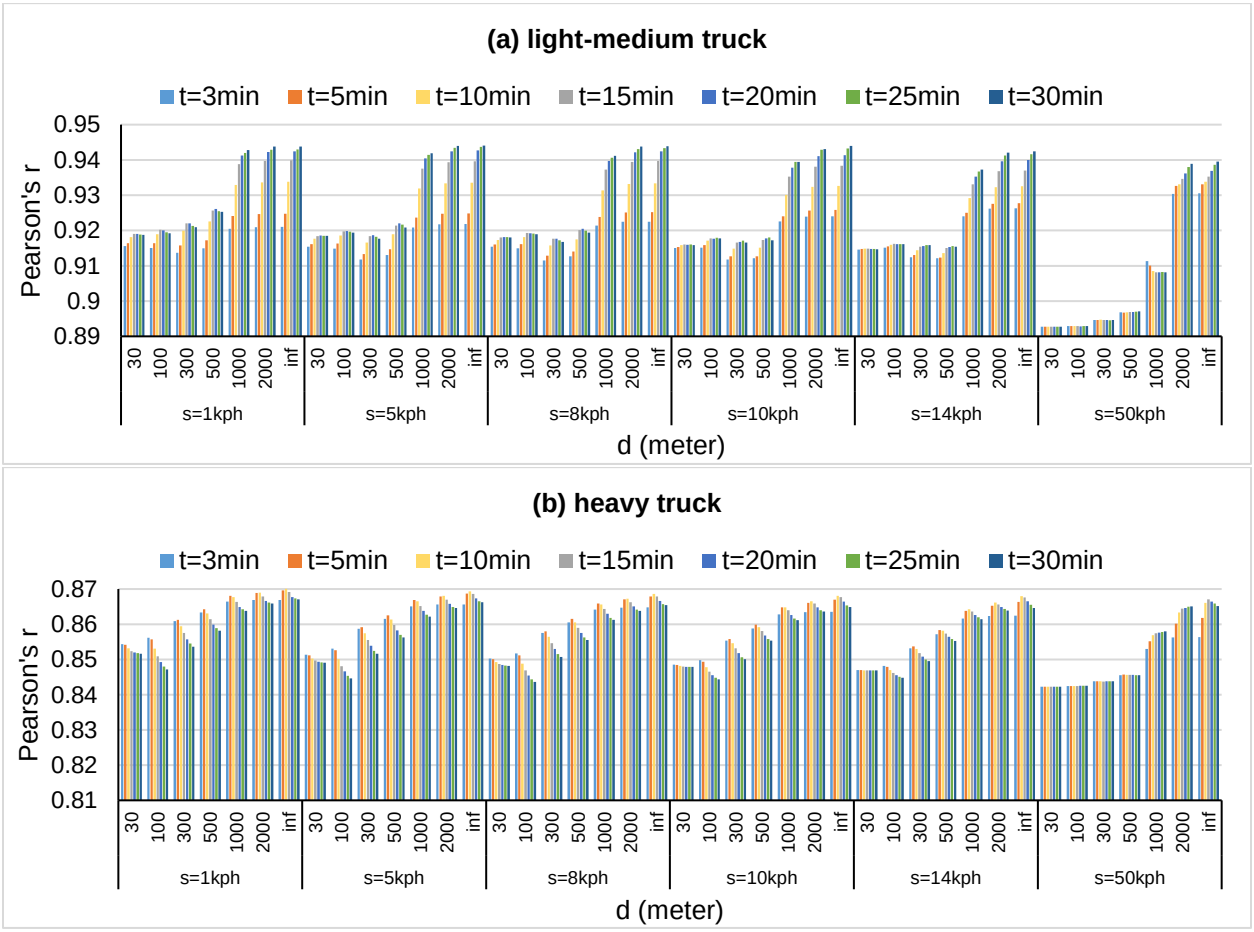


Figure 9. Sensitivity tests on threshold values.

Table 9. Discussion of threshold value test result.

Scenario	Pearson's r	Truck ratio	Trends		
			s	d	t
Figure 9a: light-medium-weight truck	0.89~0.94 (SD 0.01)	7%	Regardless d or t, as s decreases, the p-value increases.	A higher d, even positive infinite, yields much higher p-value than a lower d.	When d >= 2000, as t increases, the p-value also increases, regardless of s.
Figure 9b: heavy-weight truck	0.84~0.87 (SD 0.01)	93%	Same as above.	Same as above.	When s <= 14 and d >= 2000, t = 10 yields the highest p-value.

It is noted that the pattern of dwell time threshold is different between the two truck types and it seems to be counter-intuitive that a longer dwell time threshold is preferred for light-medium trucks for closeness to FAF distribution. The major reasons for this are: (1) FAF has a data gap of intracounty truck shipments while GPS data capture these relatively short intracounty trips. For light-medium truck results (Figure 9a), a larger dwell time threshold identifies less short trips, making the result closer to the FAF data, i.e., a higher p-value. An investigation shows that generally the higher p-value is, the fewer identified trips there are and mainly intrastate trips are reduced as shown in Figure 10. When increasing the speed threshold or decreasing the dwell time threshold, the p-value is smaller. This is consistent with the light-medium truck travel feature with a higher cruising speed and more frequent stops (shorter dwell time and smaller activity space at stop). (2) FAF's long-distance OD truck freight tonnage is reliable while long-distance truck trips may be partially retrieved when identifying trips from GPS data due to missing GPS logs. For closeness to FAF distribution, a higher dwell time threshold value is preferred at the beginning but later on a summit point of 10 min is reached (Figure 9b). Although loading/unloading activities of heavy trucks

usually require a longer time such as 30 min, heavy trucks also stop for other purposes such as fuel refill, meals, waiting at an intersection, and traffic congestion, which do not need a long dwell time. 10 min is acceptable for heavy truck trip identification. Very large and even positive infinite distance thresholds yield the highest p-value, meaning that releasing the distance limit and just dwell time limit itself is sufficient to detect trip ends. Fully relaxing the distance limit is beneficial for identifying longer trips. Overall, the p-value changes very slightly with a standard deviation of 0.01 meaning that the trip identification algorithm is reliable regarding robustness. Since light-medium tuck trip identification does not have a good ground truth data for deciding thresholds and the fraction of light-medium truck is only 7%. A uniform set of threshold values are finally applied to both truck types, which are decided based on heavy truck threshold test: s = 2 mph, d = inf, and t = 10 min. If reliable data are available in the future for deciding parameters, investigating customized parameters for each truck type is recommended.

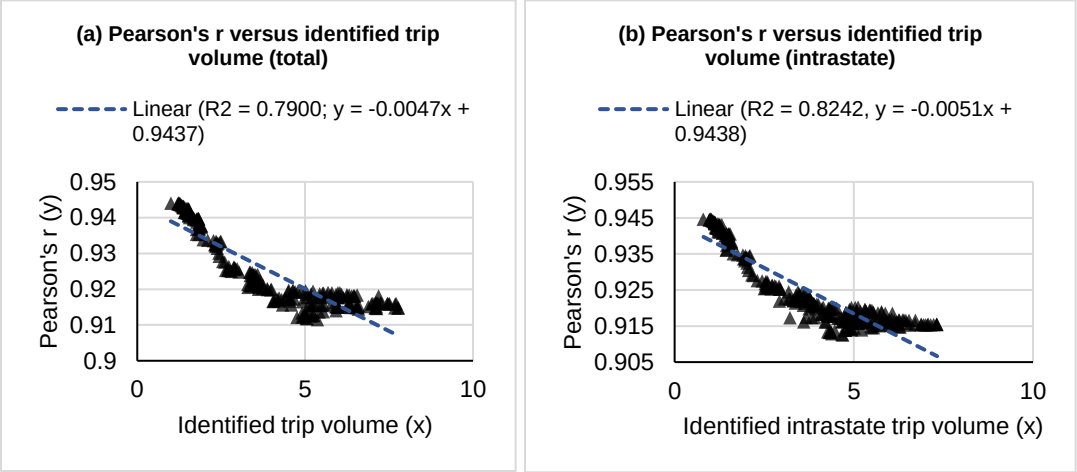


Figure 10. Pearson's r versus identified trip volume.

The truck trips identified by the threshold-based algorithm above are trips with stops from the moving status point of view. Some types of trips stop because of traffic

congestion, resting at truck rest areas, fueling at gas stations, etc., are less important than trips that stop for delivery purposes. These stop types are set as intermediate stops and trips are chained together at these places through a trip chaining algorithm, which is proposed and is peer-reviewed in a previous work co-authored by the present author (Zhang et al., 2021). The main idea is that two trips are chained at non-delivery locations using a set of truck- and trip- related features. The non-delivery locations consist of the selected places from the HERE POI dataset and highway network including truck parking lots, truck rest areas, gas stations, and major arterials. The set of features used for pairing two trips at a non-delivery location includes truck weight class, dwell time, and the differences between two trips regarding GPS point logging frequency, OD direction, and average speed. If it is dataset II, two trips with the same device ID can be directly chained together at a non-delivery location. If it is dataset I, pairing is conducted at each non-delivery location to find the trip pair with the highest possibility of coming from the same device ID. The possibility is measured by dwell time at the non-delivery location and the features of the two trips as already mentioned. Lastly, a post check is conducted to avoid over-chaining, i.e., these chained trips with a high detour factor (e.g., 1.5) after chaining are broken down into separate trips.

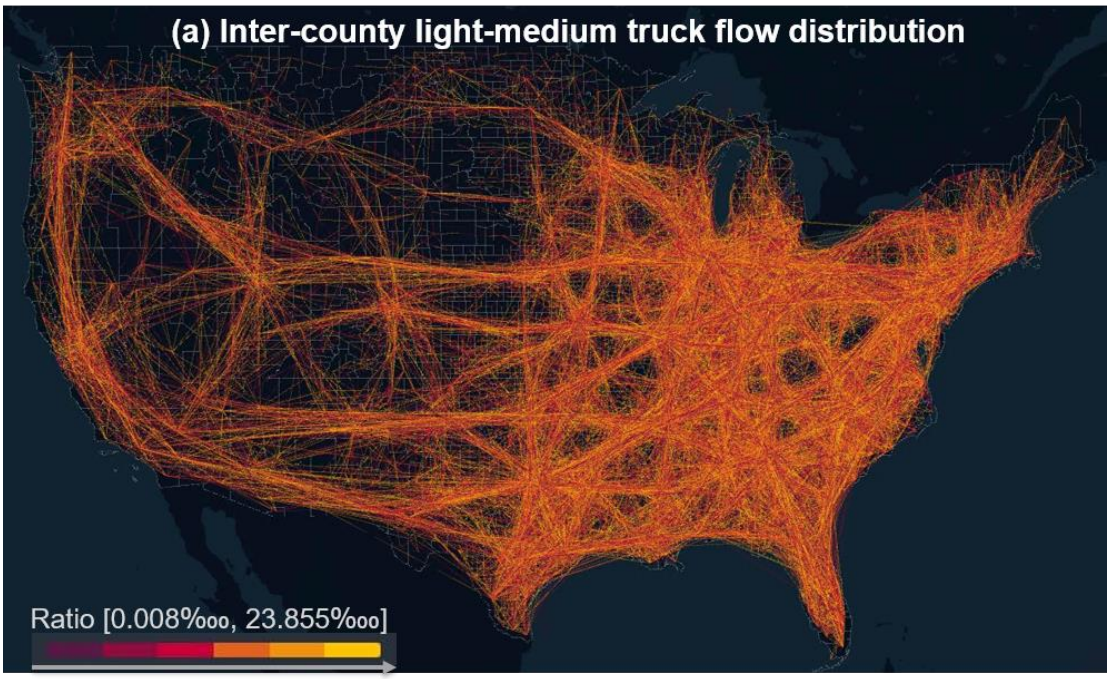
6.3. Results

Finally, the prepared trip profile contains 140 million trips for 8 million trucks. The OD truck traffic flow by truck type at county level is produced from the truck trip profile. The coverage of origins and destinations is summarized in Table 10 showing a very high level of coverage.

Table 10. Coverage of the origins and destinations by sample GPS truck flows.

Truck Type	County-level OD Flow	Origin		Destination	
		State	County	State	County
Light-medium	Inter- and Intra-County	51	3127(99.5%)	51	3127(99.5%)
	Inter-county	50(No HI)	3117(99.2%)	50(No HI)	3117(99.2%)
Heavy	Inter- and Intra-County	51	3127(99.5%)	51	3127(99.5%)
	Inter-county	50(No Hawaii)	3119(99.2%)	50(No Hawaii)	3119(99.2%)

Figure 11 shows the distribution of inter-county truck flows by truck type based on the prepared truck profile. It is obvious that heavy trucks present more long-distance truck flows. The trip identification and trip chaining algorithms successfully capture the difference in spatial pattern between the two truck types.



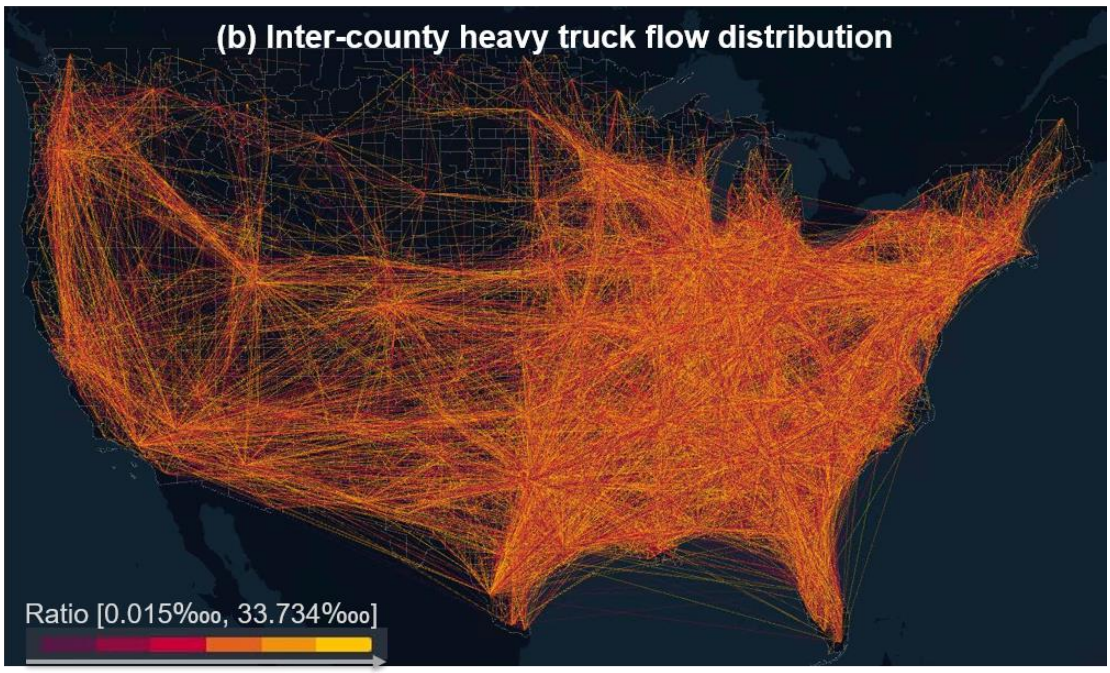


Figure 11. Inter-county truck flow distribution by truck type.

The hourly distribution of estimated GPS truck trips is compared against TMAS annual-level hourly distribution as shown in Figure 12. The corresponding p-values are 0.91 and 0.95 respectively. Generally, the identified trips from GPS data share a very similar trend to the TMAS, except two minor differences: the morning and afternoon peaks in light-medium truck GPS trips are less differentiated from each other than that of the TMAS; heavy truck GPS trips' peak around noon is higher than that of the TMAS. These differences are acceptable since a moderate but not perfect similarity is expected. There is a major difference between TMAS and truck GPS trips. The hourly distribution from TMAS is based on observed truck traffic going through sensors and the hourly distribution from the identified GPS truck trips is based on the trip start time. Despite this difference, there is no ground truth data for the unique features reflected from GPS truck trips and TMAS provide the best available data for investigating

temporal patterns. Based on this comparison, the temporal bias is not serious in the GPS data used for this study. Hence, no technique is applied to adjust the temporal bias.

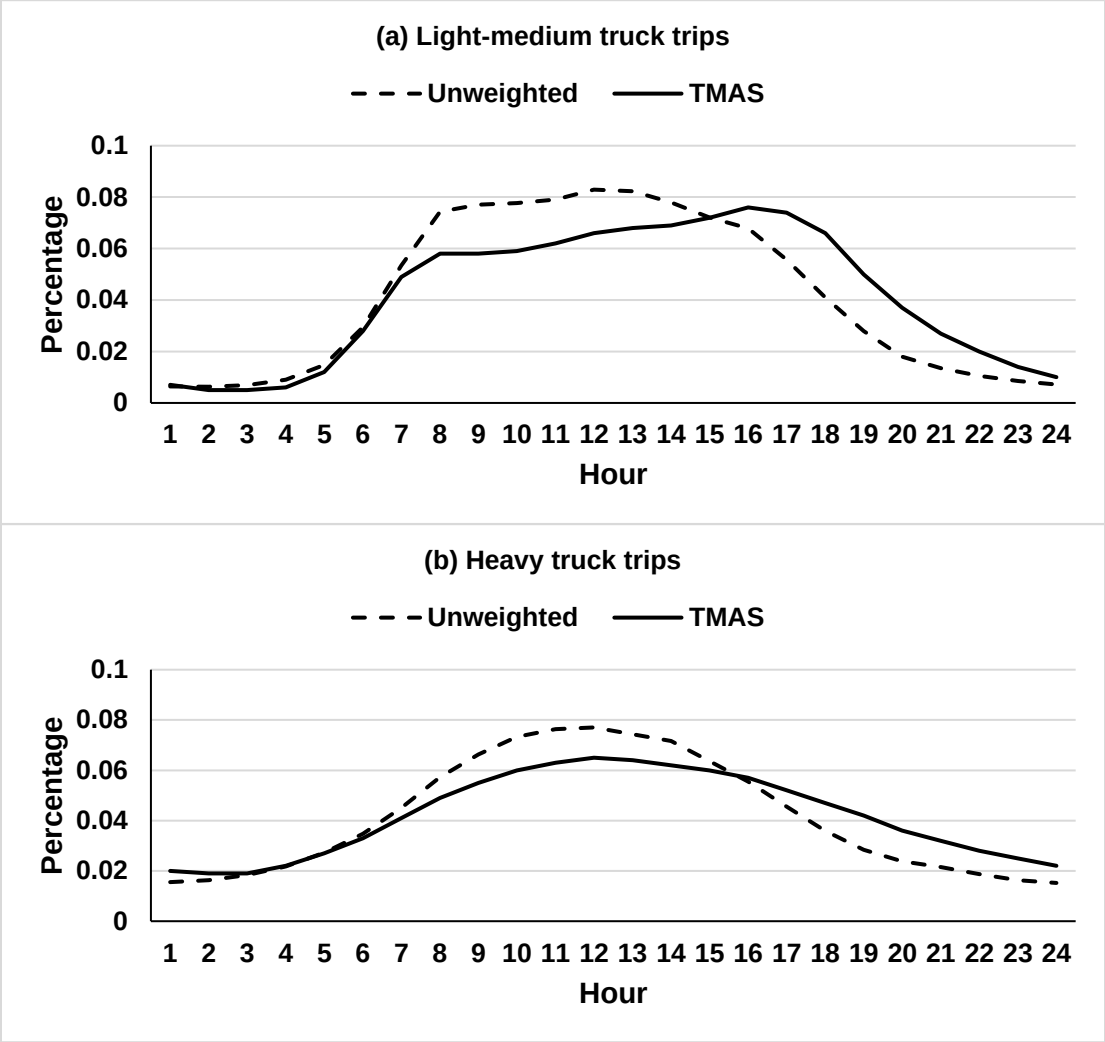


Figure 12. Hourly distribution of estimated GPS truck trip.

Chapter 7: Iterative Reinforcement-learning-based ODME

7.1. General Framework of Iterative Reinforcement-learning-based ODME

Current studies on estimating truck travel demand from GPS data mainly reply on an ODME process. In this process, GPS data is used for producing the seed OD matrix and truck sensor counts data from FHWA's Travel Monitoring Analysis System (TMAS) is used as the data input of an optimization problem. A major reason for the limited framework for truck travel estimation from GPS data is the lack of ground truth data. The truck sensor counts data are usually the only available data, the utilization of which requires an ODME process. As a recap, the current ODME process of estimating truck travel demand from GPS data includes the following four steps:

1. Preparing a seed OD matrix from truck GPS data;
2. Assigning the OD matrix to the road network with a traffic assignment model;
3. Adjusting the OD matrix by minimizing the difference between assigned traffic flows and the observed traffic counts on the road links by an optimization model;
4. Repeating the previous two steps until the difference is reduced to a desired level.

A major bottleneck for applying this process to a nationwide truck travel demand estimation is the prohibitive computation cost of traffic assignment model in step 2 and optimization model in step 3. This process is not ready-to-use for such a large area as the national road network in the United States. The traffic assignment model and optimization model are too complicated for a nationwide network. So far,

the largest scale at which this process has been applied to is Florida by Zanjani et al., 2015 and Indiana by Bernardin et al., 2011. In addition, the travel cost input for traffic assignment does not exist at national level. The optimization model requires powerful computation resources. In replace of optimization models, some ODME apply feedback-based strategies. These strategies achieve similar accuracy level without rigorous computation requirements so they have a higher feasibility. A traffic simulation model is usually integrated into the feedback-based strategies to provide knowledge for the ODME process, but the traffic simulation models usually have deficiency and indeterminacy. Hence, the current ODME framework of utilizing GPS data for truck travel estimation is not applicable to this study due to various limitations.

A novel reinforcement-learning-based ODME algorithm is designed, which presents a desirable accuracy level. It integrates the reinforcement learning with a feedback-strategy-based ODME process. There are three objectives for this algorithm: (1) to weight the trips for matching the observed truck counts collected by sensors on highway network; (2) to alleviate the spatial distribution bias; and (3) to avoid truck type bias by conducting the weighting process by truck type. Briefly, the algorithm assigns weights to the trips based on the grouping of trips by origin-destination-path. An origin-destination-path is defined as a sequence of truck count sensors that a trip goes through in a time sequence. For the trips that do not go through any sensors, different cases are considered and corresponding weighting techniques are applied. The overall weighting procedure is shown in Figure 13.

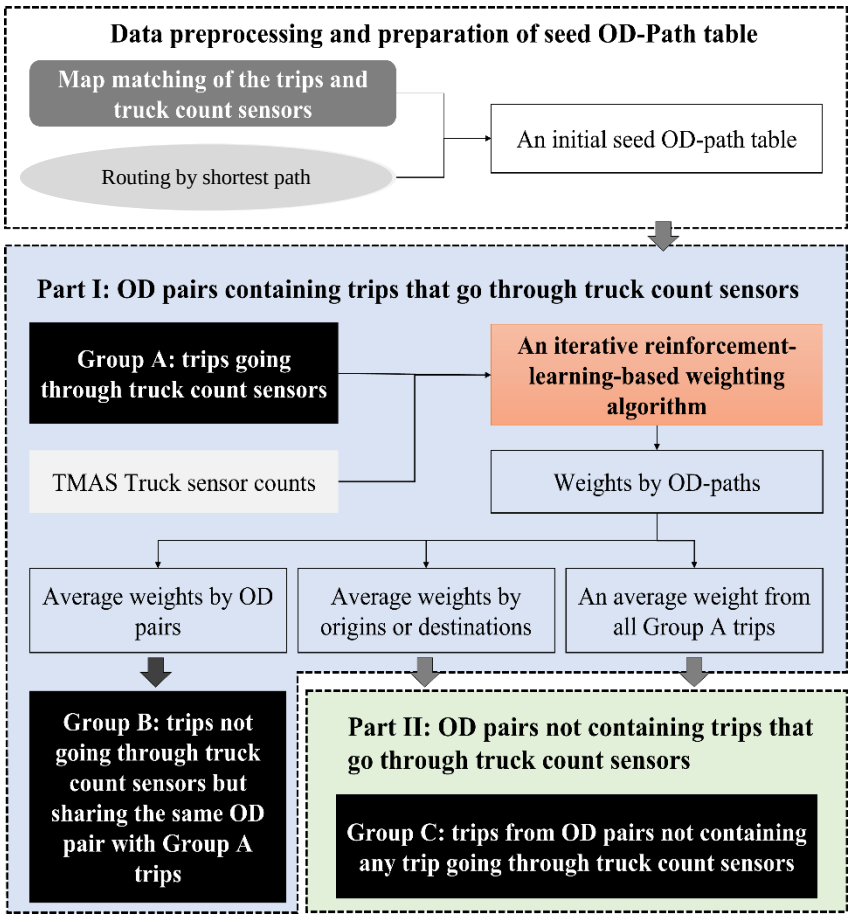


Figure 13. The structure of weighting procedure.

Data preprocessing is needed at first to do map matching and routing. Map matching is for mapping the trips and truck count sensors to the highway network. Routing is needed to fill up the missing trip segments due to unlogged GPS points in passive GPS data. Then, the OD-path information is appended to each trip. An OD-path is defined as a sequence of truck sensors that a trip goes through in time order along with the origin and destination of this trip. An initial seed OD-path table is therefore obtained, which tabulates each OD pair by paths. This initial seed OD-path table along with the observed truck counts at each sensor are the data input for the

iterative reinforcement-learning-based weighting algorithm. The output of the algorithm is the weights for all OD-path groups.

All OD pairs from the seed table are divided into two parts. Part I is the OD pairs that have trips going through truck count sensors. Part II is the OD pairs that do not have any trip going through truck count sensors. All trips from Part I OD pairs are divided into two groups, namely Group A and Group B, while all trips from Part II OD pairs are Group C trips. The weighting details for each group of trips are provided below.

Part I - OD pairs containing trips that go through truck count sensors:

- Group A consists of trips going through sensors: an iterative reinforcement-learning-based weighting algorithm is implemented. The weighting is conducted at OD-Path level meaning that the trips belonging to the same OD-path group are assigned with the same OD-path weights. The algorithm is described in detail in later sections.
- Group B consists of trips not going through sensors: for each OD pair, an average weight is derived from the corresponding weighted Group A trips and is applied to group B trips.

Part II - OD pairs not containing any trip that goes through truck count sensors:

- Group C consists of trips all from Part II OD pairs: although a Part II OD pair cannot be directly weighted through the algorithm. Its

origin or destination may be related to a Part I OD pair. In this case, an average weight based on origin or/and destination is used. The origin or destination weight is the ratio of total weighted traffic to total initial traffic of Part I OD pairs given an origin or a destination. If either origin or destination weights exist, only that origin or destination weight is applied. If both origin and destination weight exist, the squared root of the two weights is applied. If neither origin nor destination weight exist, an average weight derived from Part I OD trips, i.e. the ratio of total weighted traffic to total initial traffic of Part I OD pairs, is applied.

The percentage of OD pairs by weighting technique is summarized in Table 11. For both truck types, most of the trips, i.e. 98%, are weighted through the reinforcement-learning-based algorithm and a very small fraction of trips is weighted by part I trip weights.

Table 11. Percentage of OD pairs by weighting strategies.

Group	Weighting strategy	Light-medium	Heavy
Part I - OD pairs containing trips that go through truck count sensors	Iterative RF-based weighting	98.027%	97.596%
Part II - OD pairs not containing any trip that goes through truck count sensors	Proportional weighting based on the origin and destination weight from part I	1.970%	2.401%
	Proportional weighting based on the origin weight from part I	0.002%	0.002%
	Proportional weighting based on the destination weight from part I	0.000%	0.000%

	Proportional weighting based on the average weight from part I	0.001%	0.001%

7.2. An Iterative Reinforcement-learning-based Weighting Algorithm

7.2.1. Reinforcement Learning and Q-Learning

Traditional ODME approaches take constant adjustment rate or adjustments in each iteration. This simple algorithm is efficient in finding optimal solutions for small network with limited observations. However, a large network with rich observations is hard to benefit since the complexity and the otherness of the network usually result in the early and incomplete coverage of the optimization process. As one of the most popular learning methods for complex relationship, reinforcement learning (RL) is a potential method for addressing complex optimization problems. As one of the three basic machine learning methods, reinforcement learning is utilized to train the agent on how to taking actions under different environment to maximize the potential benefits. Different from supervised learning method that requires accurate labels and explicit sub-optimal actions, reinforcement learning focus more on finding the balance between exploration and exploitation.

A reinforcement method considers the optimization process as a Markov Chain Decision process with the assumption that the future state is completely independent of the previous states and actions, as indicated by equation (3):

$$P(S_{t+1}|S_t) = P(S_{t+1}|S_1, \dots, S_t) \quad (3)$$

where S_t refers to the current state at time t and S_{t+1} stands for the state of the next timestamp.

With this kind of relationship, the reinforcement leaning process tends to generate a set of sequential decisions under an observable environment. A standard reinforcement model consists of three parts: agent, environment, and reward function. The agent is a self-learning machine that exchanges information with the environment on states, actions, and rewards. Reward function provides the metrics to evaluate the performance of the environment under certain states and actions, which is the source for the agent to learn and improve. At each time step, the agent perceives the states of the environment and takes an action to transfer the environment from one state to another state. The actions under states are evaluated using the reward function. The state-action-reward information is recorded and updated. After several iterations, the agent traverses enough state-action pairs and learns how to find optimal solutions with maximum rewards.

Among various RL approaches, Q-learning (QL) is the most widely used in real-world implementations. In the Q-learning process, the self-learning is achieved through the Q table, which stores the state-action-reward information and is updated iteratively. Without defining a policy, the agent learns and improves through the historical information in the Q table.

7.2.2. Integration of Reinforcement Learning to ODME

Instead of applying a fixed adjustment rate to all sensors and all iterations, a dynamic rate can be determined through QL process. In the QL-based ODME strategy, the agent communicates with the environment by receiving states (i.e., weighted mean squared error between simulated and observed volume) and taking actions (i.e.,

implementing the adjustment rates). A reward function evaluates the performance (i.e., the change of weighted mean squared error) of state-action pairs. Even though the QL-based ODME strategies require training time in the offline simulation environment, the well-trained agent can take optimal actions under various states without complex computation and accurate prediction models. The objective of most ODME strategies is minimizing the difference between the observed traffic volumes and the simulated traffic volumes, which can be evaluated by the weighted mean squared error.

In the QL-based strategies, actions change the environment from one state to another. A fixed adjustment rate for all the iterations across all the sensors can be applied. However, this setup has two obvious drawbacks. Firstly, different OD pairs contribute to the system at various degrees. For instance, an OD-path crossing a single sensor can adjust this sensor more effectively compared to another OD-path crossing multiple sensors with different error rates. The fixed adjustment rate reduces the efficiency of the entire system. Secondly, some sensors may quickly reach the bound and are hard to be adjusted furthermore with the fixed adjustment rate. Therefore, this study separates the OD-paths into three groups based on the number of sensors to which they are related: (1) one sensor, (2) two or three sensors, and (3) more than three sensors. A dynamic adjustment rate (i.e., 0%, 5%, and 10%) is applied on each group of sensors during one iteration. Figure 14 shows the distribution of OD-paths versus the number of sensors involved. It is obvious that the OD-paths of heavy truck trips generally involve more sensors than those of light-medium truck trips. Very few light-medium truck trips go through more than 20 sensors. Since heavy-duty trucks also

conduct short trips under 50 miles, it is normal to have heavy truck trips going through fewer sensors.

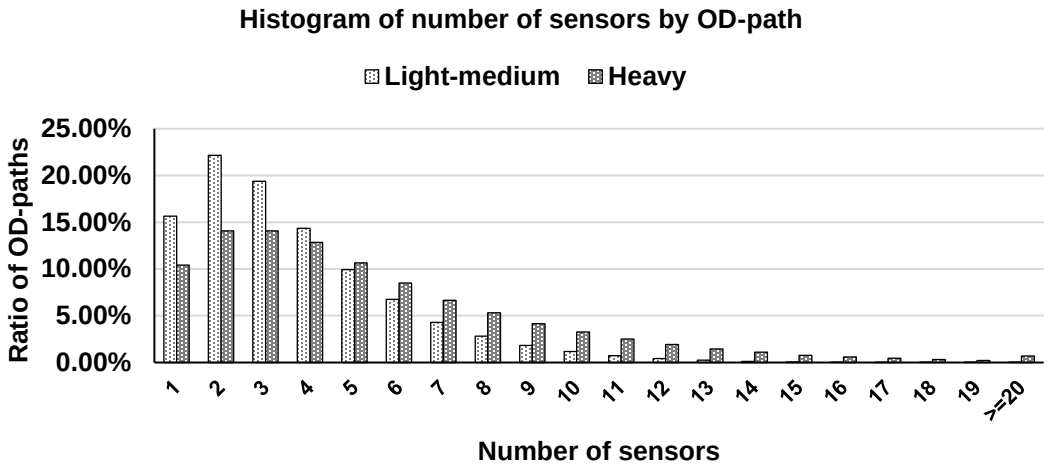


Figure 14. Histogram of number of sensors by OD-path.

The update of Q table is through the equations (4) & (5).

$$TD(s_t, a_t) = R_{s_t,a_t} + \gamma * max_a Q(s_{t+1}, a) - Q(s_t, a_t) \quad (4)$$

$$Q^{new}(s_t, a_t) = Q^{old}(s_t, a_t) + \alpha * TD(s_t, a_t) \quad (5)$$

where $TD(s_t, a_t)$ is the temporal difference for the action taken in the previous state; R_{s_t,a_t} is the reward received for the action taken in the previous state; γ is the discount factor ranging from 0 to 1 (when γ is close to 0, the agent tends to consider only immediate reward; when γ is close to 1, the agent considers a future reward with greater weight), which is set as 0.9; $max_a Q(s_{t+1}, a)$ is the largest Q-value available for any action in the current state (the largest predicted sum of future rewards); $Q(s_t, a_t)$ is the Q-value for the action taken in the previous state; $Q^{new}(s_t, a_t)$ is the

new Q-value for the action taken in the previous state; $Q^{old}(s_t, a_t)$ is the old Q-value for the action taken in the previous state; α is the learning rate ranging from 0 to 1, which is set as 0.9.

Each state-action pair is evaluated through the rewards function, which represents the environment change from one state to another state under one action. The objectives are reducing both the overall weighted mean squared error and the inequity of OD-path weights. Therefore, the reward function is formulated as equation (6).

$$R_{s_{t+1},\, a_{t+1}} = \begin{cases} -100 & if\ e_{t+1} < e_t \\ -1 & if\ e_{t+1} \geq e_t \\ 100 & if\ e_t < E \end{cases} \tag{6}$$

where R_{s_t, a_t} represents the rewards achieved by the state-action pair $s_t a_t$; e_t stands for the overall weighted mean squared error; E is the final target of weighted mean squared error, which is set as 0.15.

The iterative reinforcement-learning-based weighting algorithm adjusts the seed OD-path table by matching with the observed truck traffic from truck sensors in an iterative way. The major goal is to reduce the weighted difference between the weighted truck traffic and the observed truck counts at sensor locations to a predefined accuracy level. At each iteration and each sensor, an adjustment is made for the trips of each OD-path group going through this sensor. Adjustments are made simultaneously for all sensor locations at each iteration. Reinforcement-learning is involved by making differentiated adjustments for trips from different OD-path groups

at each iteration. Through a learning process (e.g. 200 rounds running), the best set of adjustments in time sequence is learnt by the algorithm to yield the best reward (i.e. the smallest weighted difference). This iterative reinforcement-learning-based weighting algorithm works as follows:

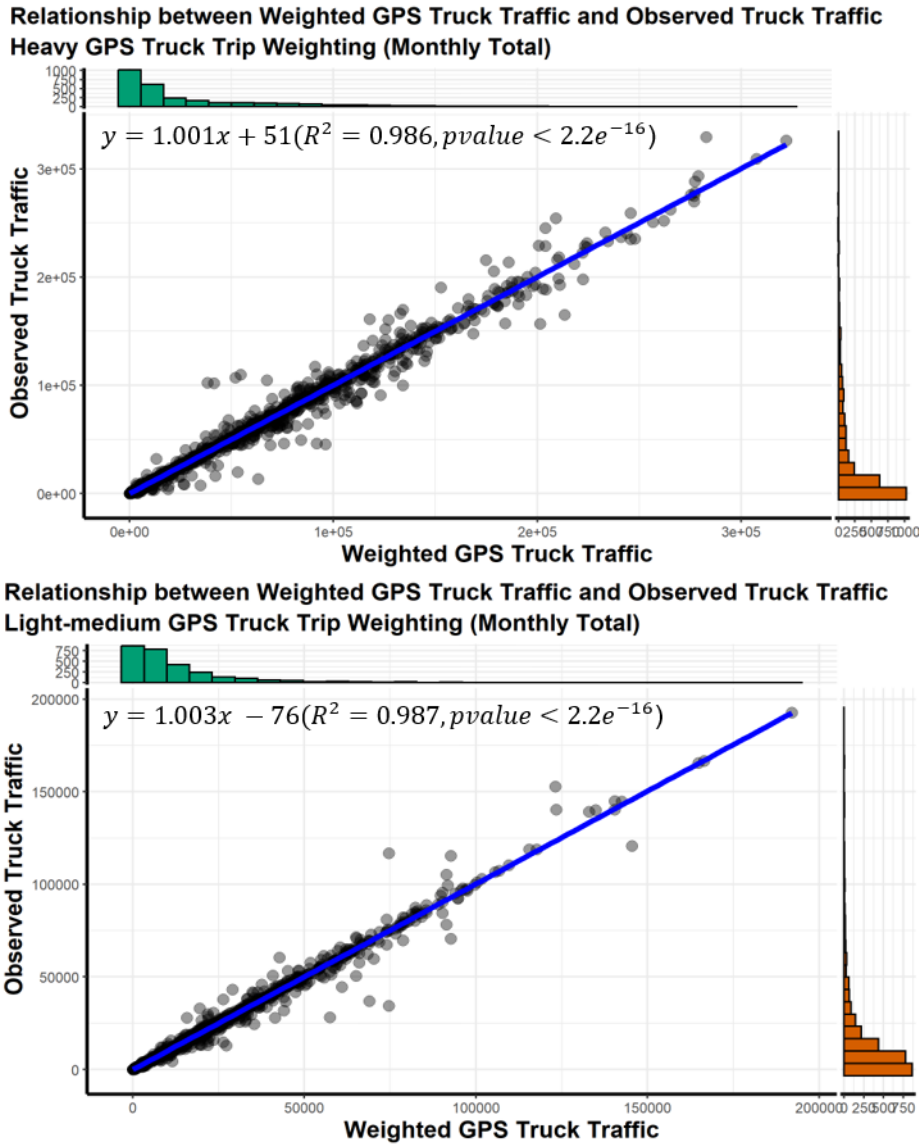
- Step 0: An initial expansion factor, i.e. to the ratio of total observed truck counts to total sample truck traffic going through sensors (total Group A trips), is applied to the initial seed OD-path table for saving computation time.
- Step 1: An iterative reinforcement-learning-based algorithm is applied at OD-path level. At each iteration and each sensor, the difference between the sum of adjusted OD-path traffic and the observed truck sensor counts is calculated. This difference is proportionally distributed to all OD-paths going through this sensor. The proportion is the initial sample traffic distribution of all OD-paths at this sensor. Each OD-path may receive such assigned adjustments from multiple sensors, the mean of which is temporarily set the adjustment for each OD-path. For each group of OD-paths, a step size (a percentage of the adjustment) is selected from an action set, which includes several options (e.g., 0, 5%, and 10%). Hence, the step size is different across OD-path groups at each iteration. Meanwhile, a marginal control of +/- 10% is included to avoid large adjustments at each iteration.
- Step 3: check if the iteration reaches a predefined accuracy level (e.g., 10%). If not, repeat the previous two steps.
- Step 4: repeat Step 0 – Step 3 for many rounds, e.g. 200, to let the agent fully explore the environment. Then an optimal set of actions (step size for each

sensor at each iteration) is learnt by the algorithm and the smallest weighted mean squared error across all sensors is obtained.

Chapter 8: Nationwide OD Truck Flow Estimation Results

8.1. ODME Accuracy Measures

The weighted truck traffic and observed truck counts are compared at sensor level as shown in the Figure 15. The scatter plots show that the weighted truck traffic volumes are very close to the truck traffic observed by sensors. The weighting algorithm works well to match with the observed truck traffic.



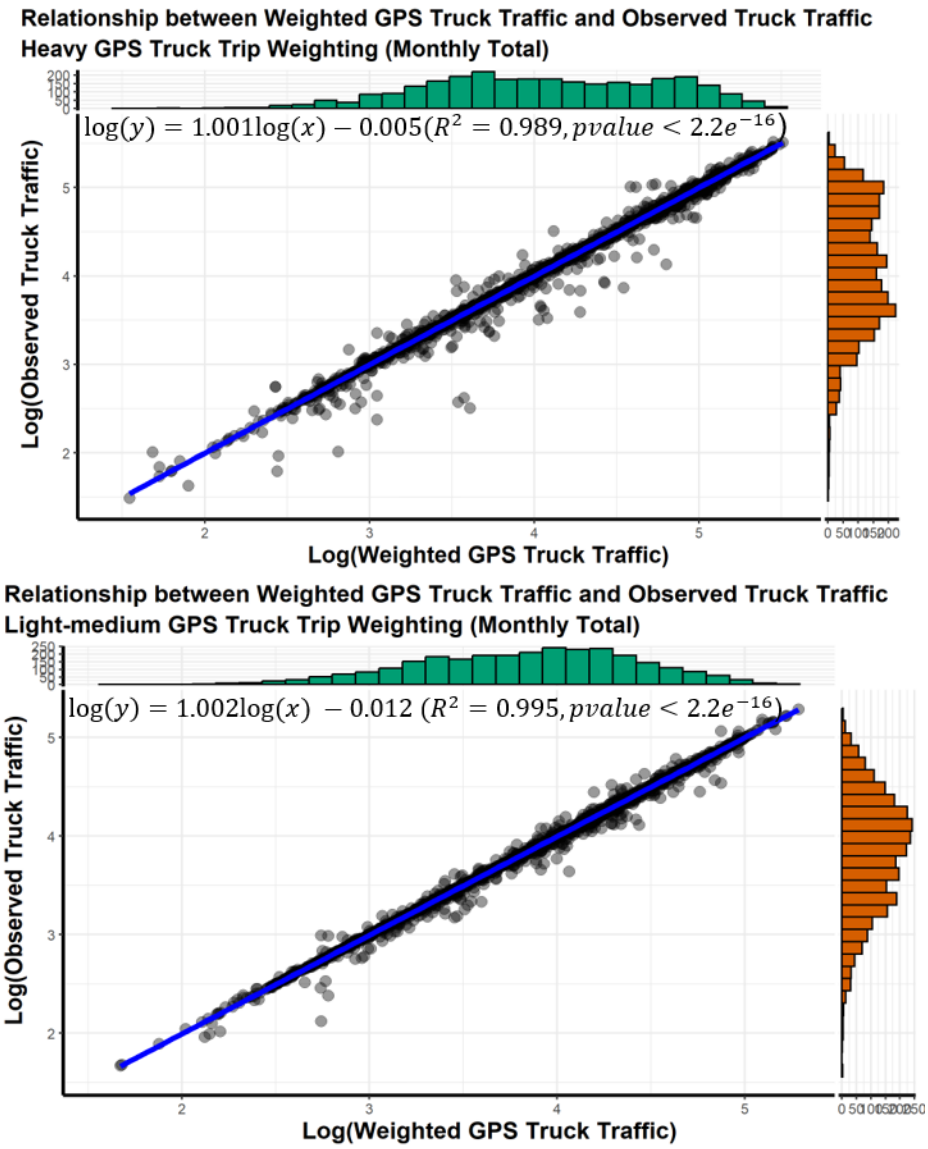


Figure 15. Weighted truck traffic versus observed truck traffic by sensor.

Figure 16 shows the average percentage error for each sensor group regarding observed truck traffic volume. For both light-medium and heavy truck traffic estimation, most sensors present no more than 0.01 million monthly total observed truck traffic volume. The average percentage error for this group of sensors is 4% and 12% for light-medium and heavy truck trip weighting respectively. This is a desirable accuracy level.

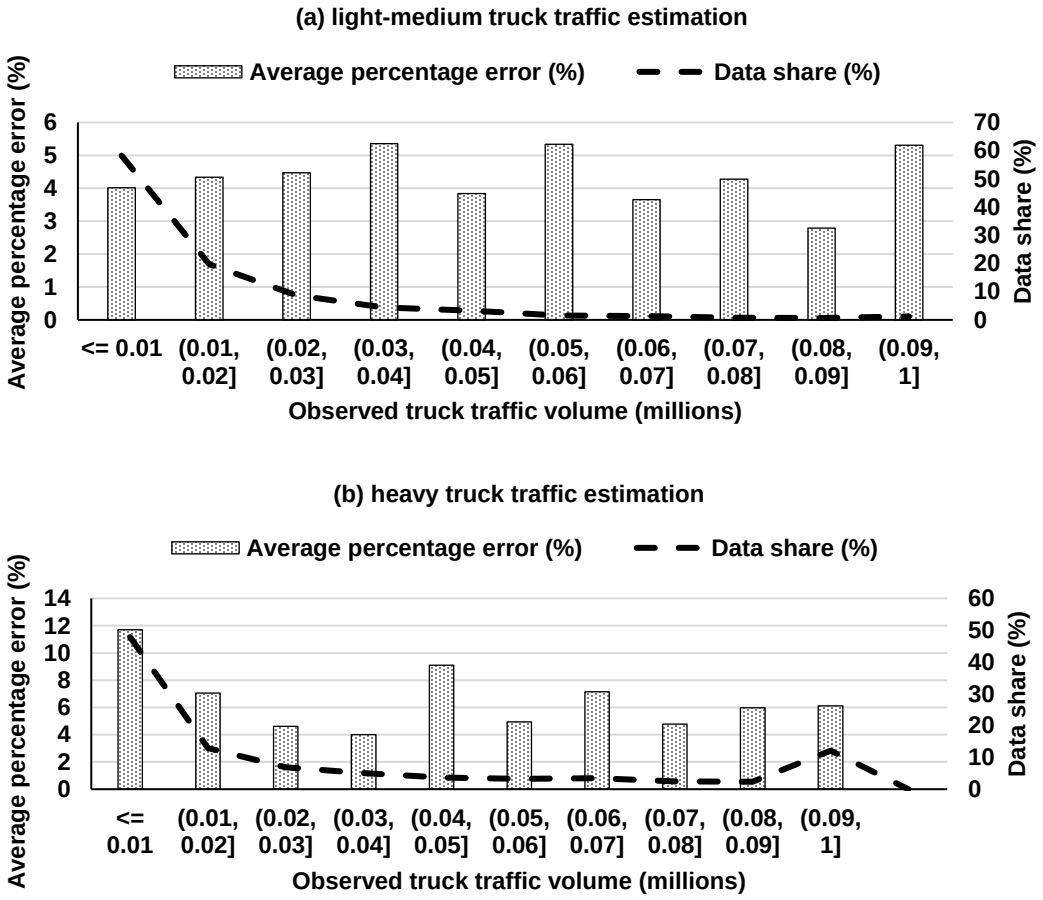


Figure 16. Error distribution across sensor volume groups.

8.2. Discussion on Sample Biases

In addition to accuracy measure, the spatial bias before and after the weighting process is measured. For each state, two ratios are calculated: (1) the sum of observed truck traffic volumes across sensors to the sum of initial sample GPS truck traffic volumes across sensors; (2) the sum of observed truck flows across sensors to the sum of weighted GPS truck flows across sensors. These two ratios are displayed in Figure 17. The ratio (1) in both scenarios show that spatial bias exists. Although many states share similar ratios (1), this ratio of some states, such as Alaska and Hawaii, is clearly higher than for other states, indicating a relatively lower sampling rate in these states.

After the weighting procedure, both scenarios show that all states share a very similar ratio (2), which is close to 1 with a standard deviation of 0.01 and 0.02, respectively. Overall, the weighting process alleviates the spatial bias.

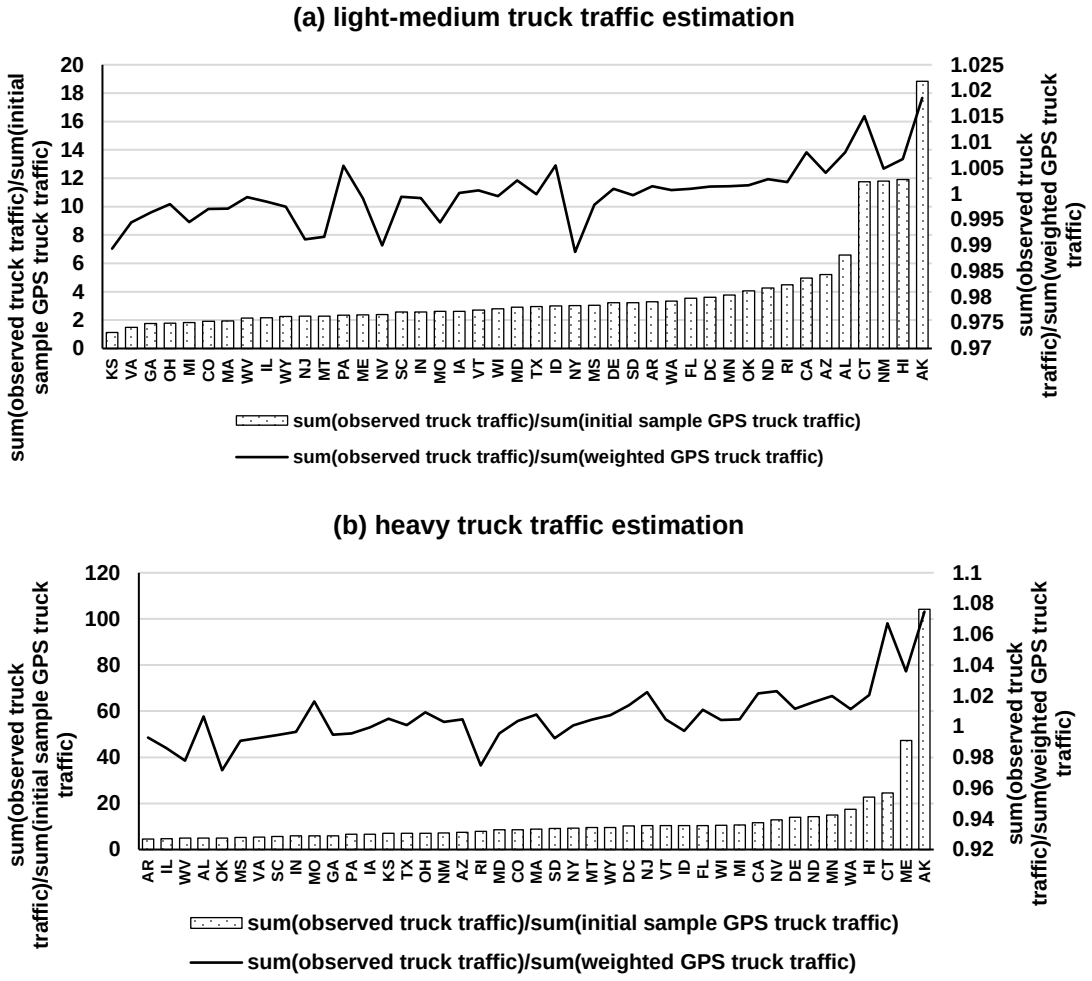
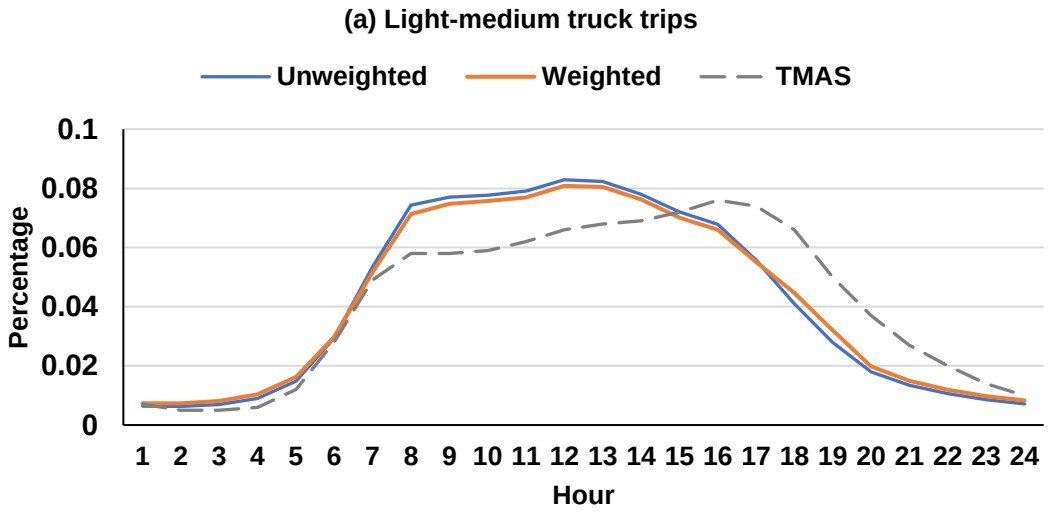


Figure 17. Spatial bias before and after ODME.

The hourly distribution if compared to TMAS sensor data. Since before the weighting procedure, the hourly distribution is investigated through the comparison with TMAS, temporal bias is not serious and no treatment is conducted to correct temporal bias. After the weighting process, the weighted hourly distribution is compared to the unweighted distribution and TMAS distribution as shown in Figure

18. First, the weighting process does not change the hourly distribution of the trips derived from the GPS. Second, the similarity to the TMAS is acceptable. As already mentioned, there is a difference between the two datasets. The TMAS hourly distribution is obtained from the through traffic observed by sensors installed on the highway system, while the GPS hourly distribution is based on trip start time. There should be some discrepancy to some extent. Overall, the correlation coefficients are 0.92 and 0.96 for light-medium and heavy scenarios respectively. There is no ground truth data for the unique features reflected from the GPS-derived truck trips. TMAS is the best available data source for temporal pattern comparison.



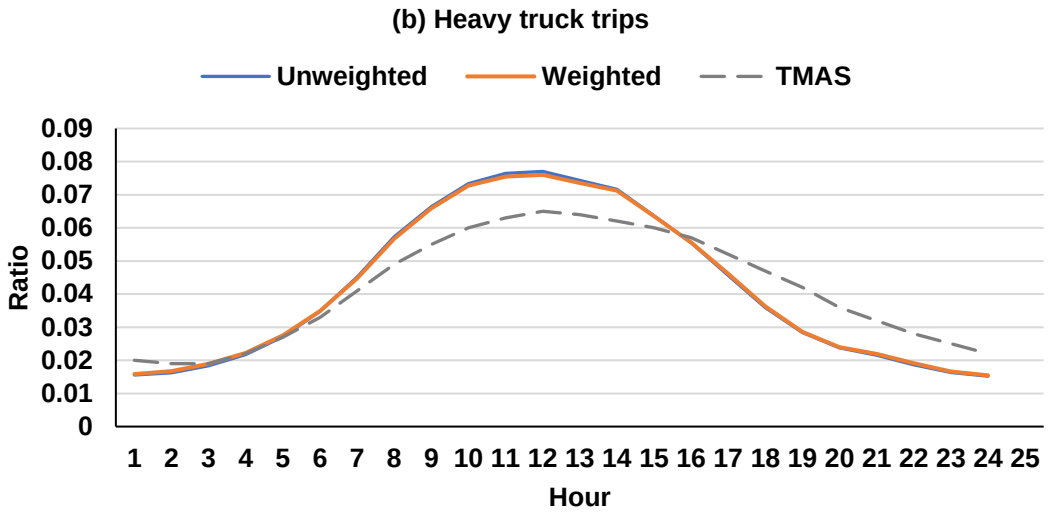


Figure 18. Hourly distribution.

8.3. Comparison to the Freight Analysis Framework OD Tonnage

The Freight Analysis Framework provides OD commodity flow by truck for the entire nation. 2020 FAF data are used for the comparison. It should be noted that 2020 FAF is an estimated product based on 2017 Commodity Flow Survey and some other data from agriculture, construction, and other sectors. A major issue with FAF is that it includes intercounty freight movements by truck mode while intracounty trucking movements are missing. Intracounty trucking has a very large share of truck movements, especially those under 50 miles. Additionally, FAF OD provides freight tonnage flow by truck. As previously discussed, freight OD tonnage flow is very different from truck traffic OD flow. Despite all these issues, FAF OD is the best available data source for nationwide OD-level information related to truck.

The Pearson coefficient between final OD truck traffic derived from GPS data and FAF OD truck freight tonnage is 0.85 at 99.9% significance level. For a better comparison, the FAF OD tonnage table is divided into two parts through the weighted

mean distance of shipments. These OD shipments below 100 miles are set as the short-haul part and those 100 miles or above are set as the long-haul part, which are compared to light-medium weight OD truck flows and heavy weight OD truck flows, respectively. The p-value in the scenario of light-medium truck is 0.76 while the p-value in the scenario of heavy truck is 0.95. The low correlation of 0.76 is very reasonable since the intracounty truck movements are not a data gap in the OD truck traffic flow extracted from GPS data. Instead, GPS-based truck traffic flow captures a large portion of local trucking, which is demonstrated by the distance distribution in section 8.4.

8.4. Distance Distribution

The truck trip distance distribution is shown in Figure 19. Light-medium truck trips are mainly less than 30 miles. For heavy truck trips, long-distance trips exceeding 100 miles are captured with (100, 300] as the locally highest distance bin. There is a clear difference between the two truck types for trips of longer than 30 miles, indicating that the proposed framework captures the different travel features by truck type. Heavy truck trips also have a high ratio of trips under 30 miles. It is normal for heavy truck trips to have many short-distance trips. First, heavy trucks (>26000lbs), such as garbage trucks, livestock transporting trucks, and street sweepers, produce short trips as well. Second, trucks are not always fully loaded. When they are empty or partly loaded, the corresponding trips are usually short.

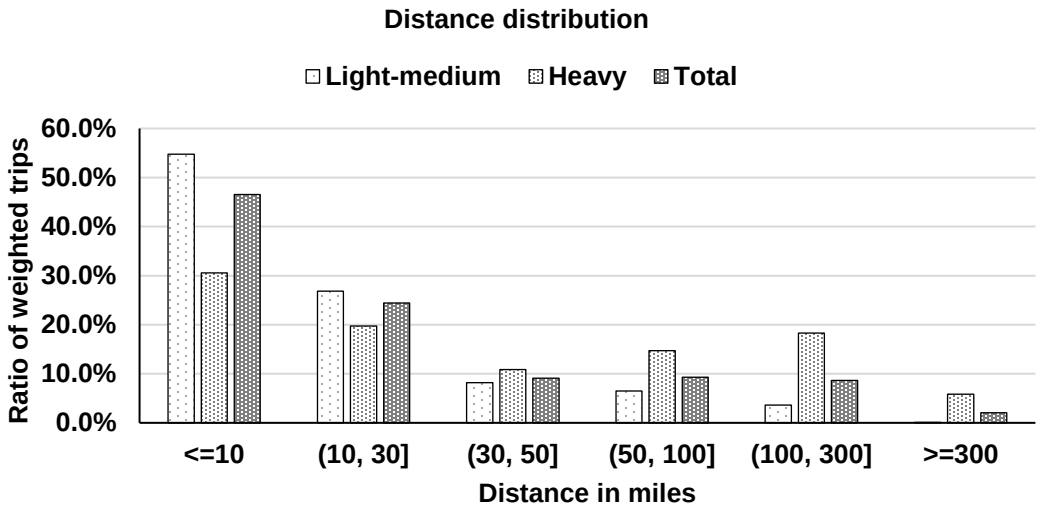


Figure 19. Distance distribution.

8.5. Vehicle Miles Traveled Validation

The VMT of the estimated truck traffic from GPS data is compared against the VMT in in Highway Statistics 2020 from FHWA based on Tables VM-2 and VM-4, as shown in Figure 20. The total of single unit and combination truck VMT are used for comparison. There are two differences between the two datasets. First, the single unit trucks contain vehicle class 4-7 while light-medium weight trucks in this study correspond to class 5-6; the combination trucks are vehicle class 8 and above while heavy weight trucks in this study correspond to class 7 and above. Hence, separate comparison by truck type is not conducted, and instead the total VMT is used for the comparison. Second, VMT from FHWA is based on road network for each state while the estimated result is based on the trip origin. For a state with a higher inter-state trip ratio, the difference between the two VMT computation methods is larger. Third, VMT from FHWA is limited to the federal-aid highway system while the passive GPS data in this study, as least not specified by data providers, are not limited to highway system. It is highly possible that light truck GPS trips occur on local roads, which are not

included in federal-aid system. Despite these differences, VMT from FHWA is the best available dataset for validating VMT. Overall, the two share a similar spatial trend with a strong Pearson correlation of 0.91 at 99.9% significance level.

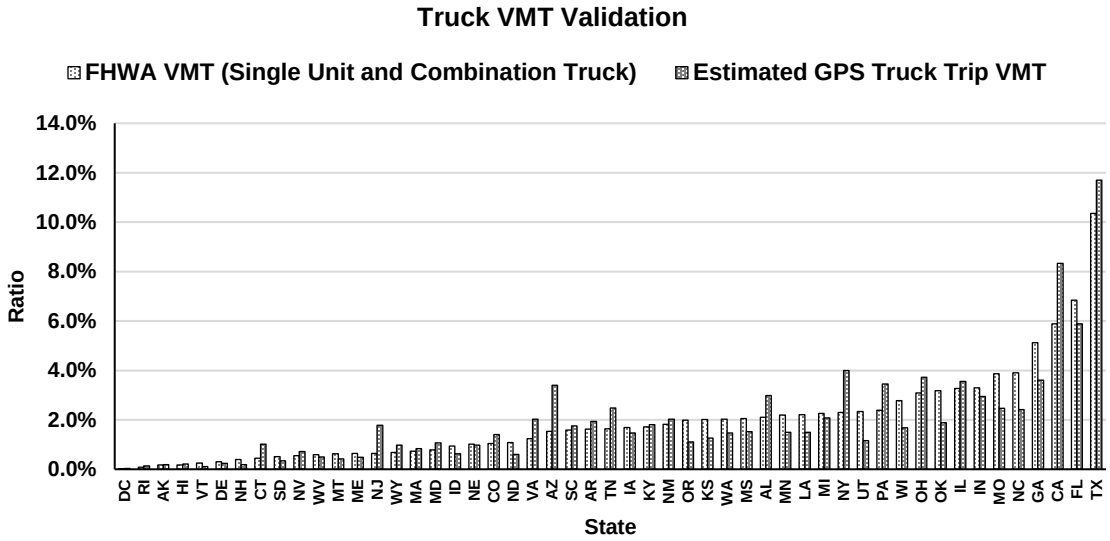


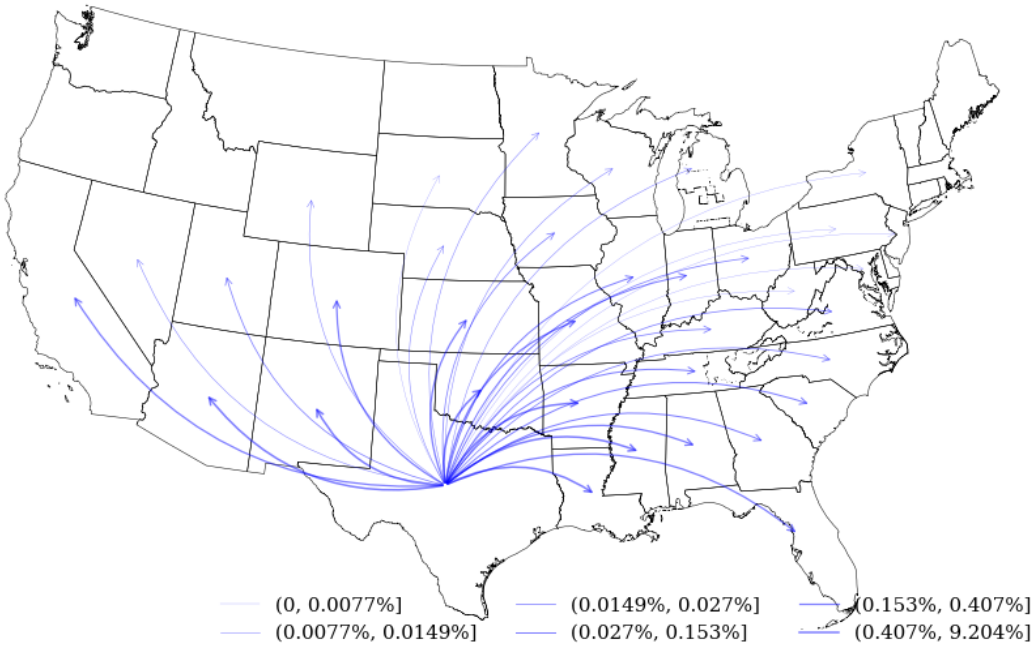
Figure 20. VMT comparison by state.

8.6. Discussion on Interstate OD Pairs from Selected States

Texas, California, Florida, and New York are the states with heavy truck activities. The distribution of interstate truck trip volume by truck type and by destinations for these states is displayed in Figures 21 - 24. OD pairs with no more than one trip on average daily basis are believed to be unrepresentative enough and are excluded. For each state, the percentage of interstate trips by each OD is shown. The summation of OD percentages in the two subfigures equal to one for each one of Figures 21 –24. The four figures present: how well long-distance trips are retrieved by the proposed framework, the spatial distribution of destination states from each origin state, the truck traffic volume share by each destination, and the difference between light-medium weight trucks and heavy trucks regarding interstate trip ratio and other

aspects. First, all four figures show that long-distance trips going across states are successfully identified by the proposed nationwide truck flow estimation framework. This is especially obvious in recovered heavy truck trips. For example, truck flow from California to Maine, from Florida to Washington, from New York to Washington, from Texas to Maine are all retrieved. Light-medium weight trucks also conduct long-distance trips across states, which are not just limited to neighboring states. Heavy truck trips present a high coverage of destination states. On average, these four origin states have 46 destination states while light-medium truck trips have an average of 25 destination states. For both truck types, closer destination states usually have higher truck flow shares. However, some states show strong attraction even though they are very far away. For example, Pennsylvania, Ohio, Georgia, and Florida are very far away from California, but they all attract moderate truck traffic from California. In comparison to other three states, Texas, located in the south of the United States, presents widely spread truck traffic flow for both truck types. Given an origin state, the traffic flow share by each destination is also easily obtained from the derived OD table. The ratio of total heavy interstate truck trips to total light-medium interstate truck trips is demonstrated by the width ratio between the two subfigures of each one of Figures 21-24. Clearly, heavy trucks have a much higher percentage of interstate trips.

(a) Light-medium: interstate truck OD flow from Texas



(b) Heavy: interstate truck OD flow from Texas

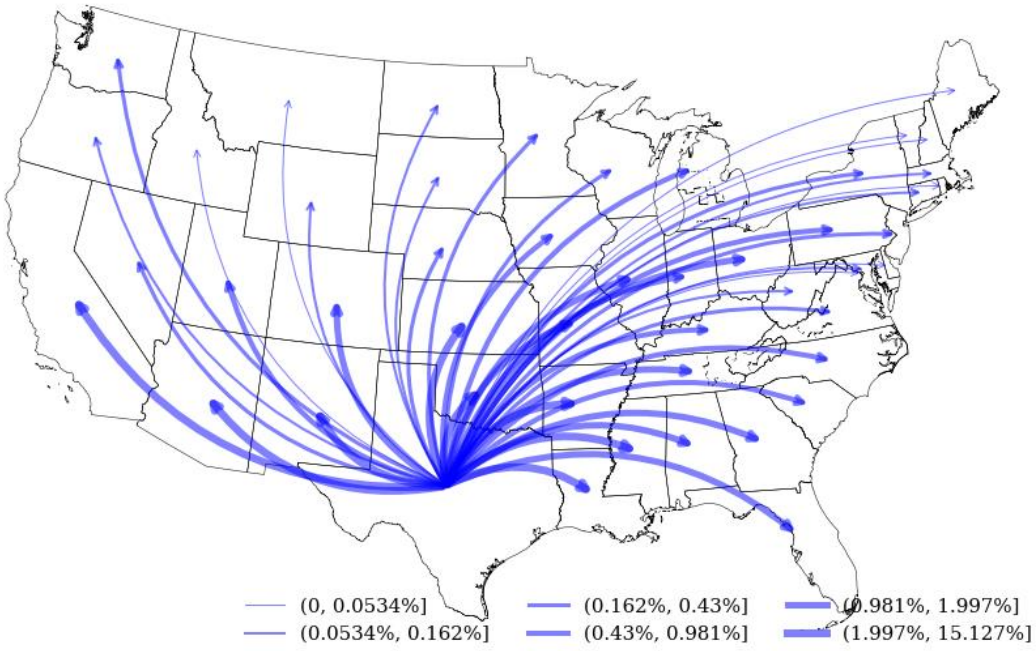
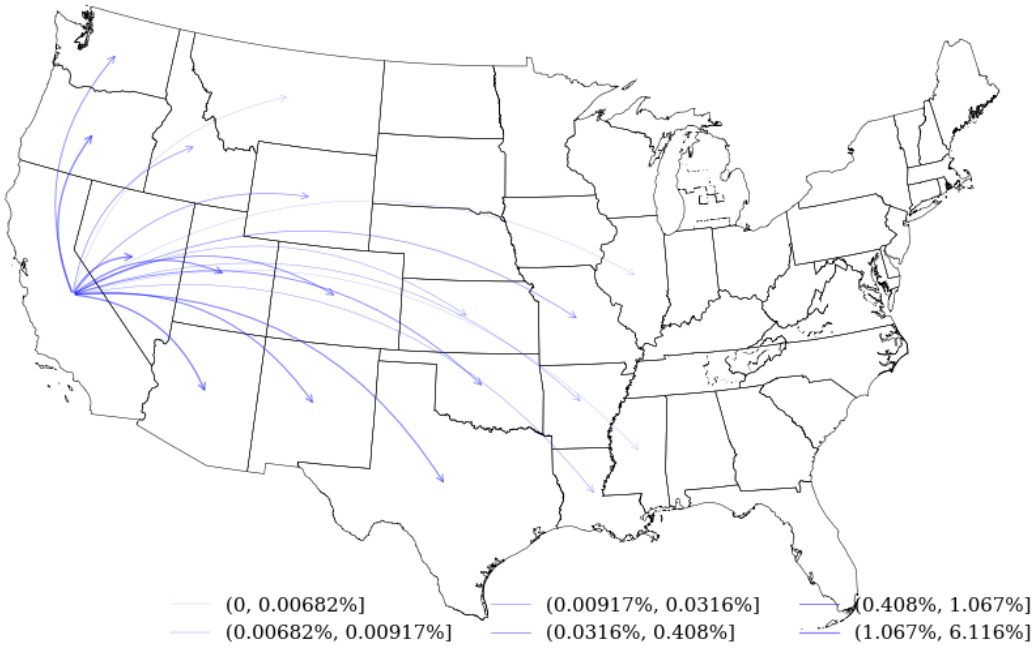


Figure 21. Distribution of the interstate truck OD flow from Texas.

(a) Light-medium: interstate truck OD flow from California



(b) Heavy: interstate truck OD flow from California

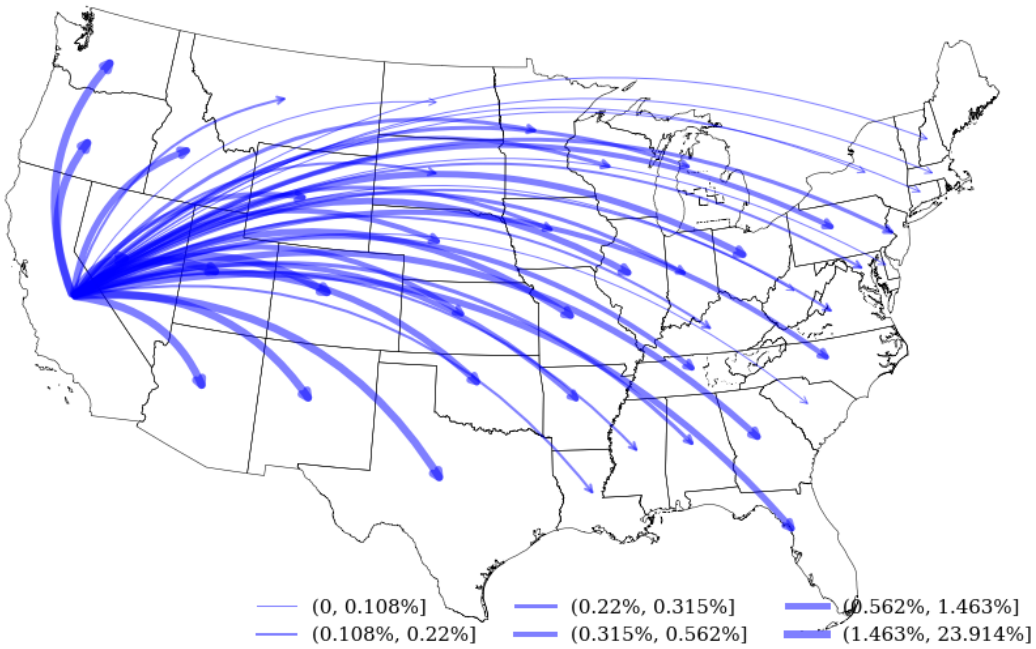
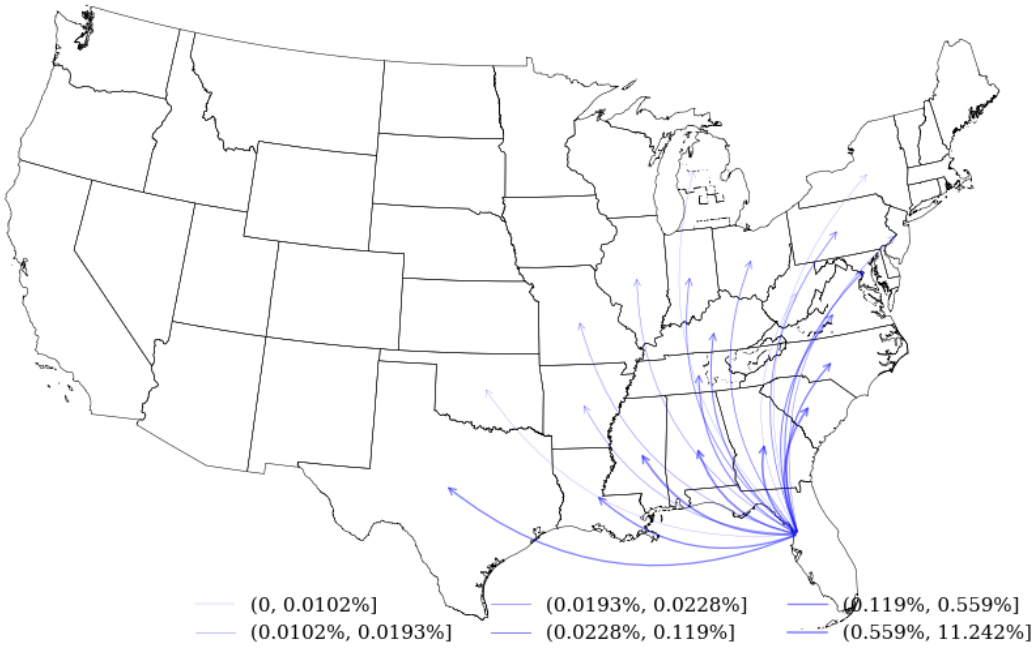


Figure 22. Distribution of the interstate truck OD flow from California.

(a) Light-medium: interstate truck OD flow from Florida



(b) Heavy: interstate truck OD flow from Florida

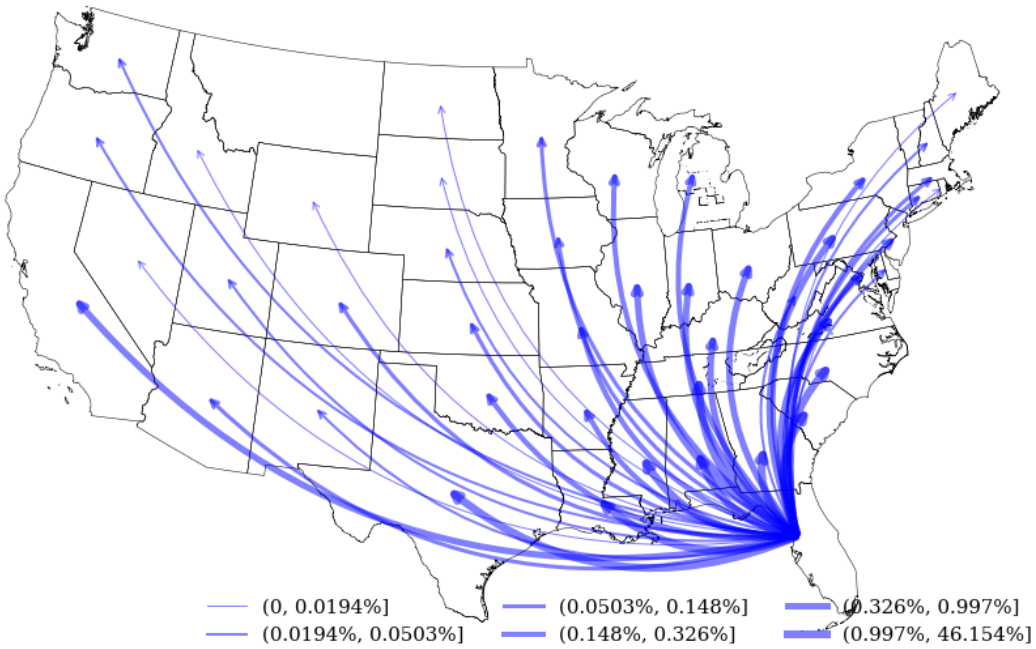
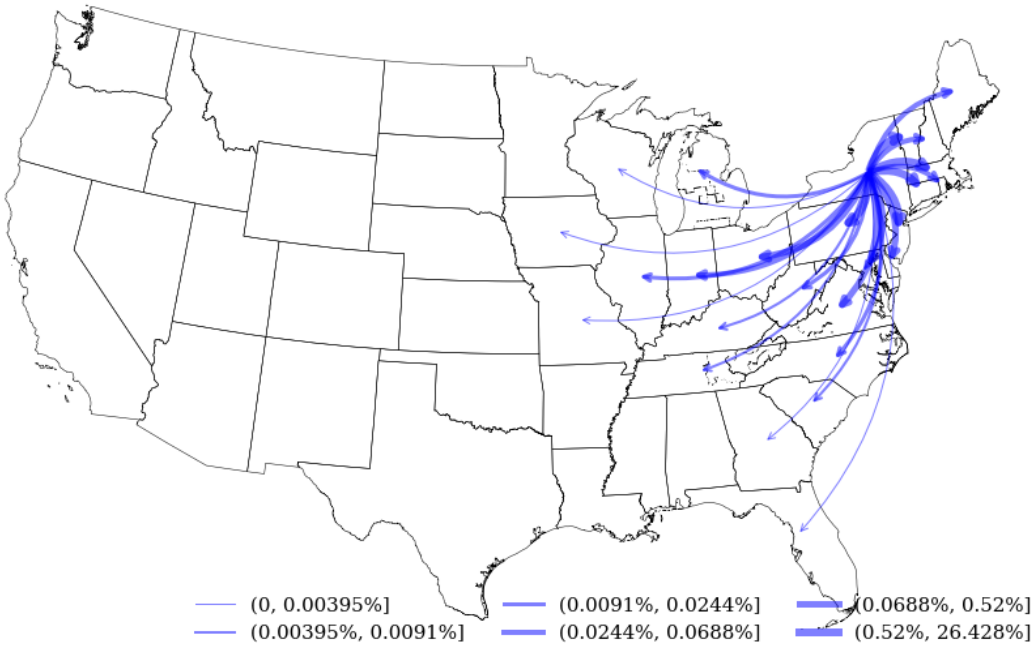


Figure 23. Distribution of the interstate truck OD flow from Florida.

(a) Light-medium: interstate truck OD flow from New York



(b) Heavy: interstate truck OD flow from New York

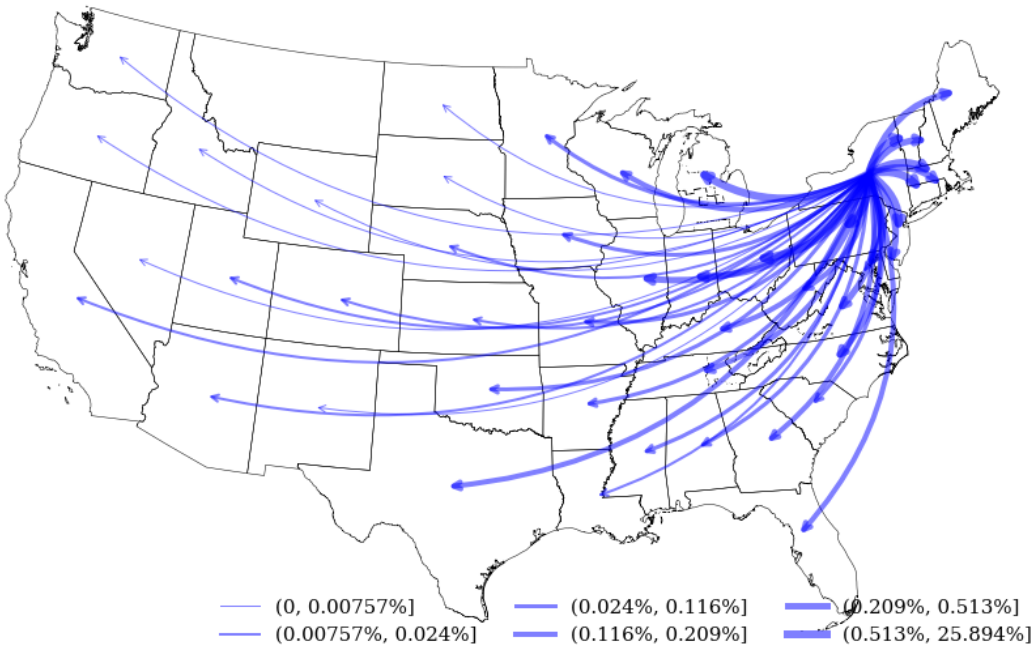


Figure 24. Distribution of the interstate truck OD flow from New York State.

Chapter 9: Conclusion

This study develops a comprehensive framework for large-scale truck flow estimation based on GPS data, with the entire United States as the case study. Although there have been some studies on GPS-based truck flow estimation, they are developed for relatively small study scales, such as statewide and citywide, and are inapplicable for nationwide estimation. To the knowledge of the author, the OD truck flow estimation at a scale as large as the entire United States is achieved for the first time through the framework developed in this study. The framework successfully fills the data gap of national truck flows in the United States, which are fundamental data bases for various transportation research areas, including operation and planning, economics, safety, logistics, environment impacts, and enhancing the efficiency, economy, safety, and sustainability of transportation systems.

Planning and operation necessitate the monitoring of truck travel on a regular basis to timely diagnose and prevent inadequacy and inefficiency in transportation systems. Particularly, more attention to be given to heavy trucks, which significantly affect highway congestion. Although sensors installed in highways capture truck flows, they are not implemented everywhere and are unable to provide origin and destination information. In contrast, the developed framework yields complete truck flows, which are a powerful supplement to sensor counts data. Truck flows also undoubtedly play a significant role in transportation safety. Special highway designs are required for accommodating trucks. Trucks increase the risk of highway accidents due to physical features of trucks and possible drowsy driving. Truck accidents tend to cause

disproportionately large losses to lives and property. For locations with high and especially hazardous truck traffic, some ways of separating truck and passenger traffic can be implemented, such as dedicated lanes for trucks and time-of-day restrictions. Truck inspection and repair services should also be helpful in promoting truck safety. From an environmental perspective, trucks consume large amounts of energy and cause substantial noise and greenhouse air pollutants. Truck flows indicate the areas that should be given adequate attention, thus helping to guide government agencies in appropriately deploying emission regulations and standards, such as setting emission limits, carbon pricing, incentive-based trucking, and Green Logistic. Decision makers should devote proper investment to alternative fuels and new vehicle technologies such as aerodynamic designs and hybrid engines. Thus, the products obtainable from the framework developed here are useful for various studies and activities.

Particularly, in all these aforementioned aspects, heavy trucks play a more critical role than light trucks. National truck flow estimation captures interstate long-distance truck flows, which are mainly due to heavy trucks. The statistics and patterns of long-distance travel are valuable but difficult to be derived from statewide or local analysis. Nationwide truck flow estimation is an asset for assisting national and regional decision-making processes. It provides a comprehensive perspective for promoting coordination and cooperation between states. As exemplified by national OD truck flows for California, Texas, Florida, and New York, a large fraction of trucks travel across several states, but these flows cannot be captured by the statewide analysis of intermediate states. Additionally, nationwide estimation promotes the modal shift between truck and rail, which may yield safety, energy and environmental benefits.

Long-distance truck travel impacts can be mitigated by shifting some of the demand to railways.

With the United States as the case study, the results provide valuable insights into the different patterns shown by light-medium and heavy trucks in their distance and OD distributions. Additionally, the national spatial and hourly distributions of truck traffic and the vehicle miles traveled by state are informative by particularly showing the unique empirical references from GPS-based truck flow.

The framework is developed based on passive truck GPS data since utilizing GPS data in substitution for surveys demonstrates many benefits, such as being cost-effective without a huge investment in human effort and resources, being up-to-date, reduced impact of human errors, high temporal resolution to serve different temporal analyses (e.g., hourly, daily, and monthly), high spatial resolution to fill dynamic zone systems (e.g., census block group and traffic analysis zone), and ability to support special event analysis (e.g., hurricanes). The developed framework not only enables large-scale truck flow estimation from GPS data but also has strengths in completeness, thoroughness, practicability, and high-generalizability. It incorporates all fundamental research tasks including data preprocessing, vehicle type classification, truck trip identification and chaining, and weighting by truck type. With a comprehensive literature review, the algorithms involved in the framework either make improvements to existing methods or are innovatively designed to fill research gap. There are challenges and difficulties in the application of passive GPS data, which are thoroughly handled by this study. The framework is also promising in practical feasibility, since it

is straightforward and shows a desirable accuracy level. It shows high generalization ability since it is applicable to various study scales or settings. Regional and statewide studies in different zone systems or for corridors of interest are also achievable with this framework. Some other statistics, such as truck vehicle miles traveled, truck average daily traffic, travel distance distribution, travel time distribution, destination choice distribution, are all byproducts of the developed framework.

Chapter 10: Future Work

Although overall the developed framework has many strengths, it could be further improved as follows.

In data preprocessing, GPS data from two data providers are merged for supporting this research. In practice, studies based on multi-sourced GPS data are very common. When the overall sampling rate reaches very high or similarity presents between different sources, it is important to measure the duplication issue, i.e., one device is repeatedly captured by each data source. This should be a negligible issue in this study since the two datasets are dominated by different truck types, but for the completeness of framework it is desirable to discuss this and provide practical guidance. For the data oscillation identification algorithm, abnormal movement patterns are captured and demonstrated, but the underlying reasons are not well-understood. The ping-pong phenomenon along with cellular tower data is easily understood while the reasons for the other oscillation patterns are not investigated. The algorithm identifies and removes 2.4% data oscillations, which is a relatively small fraction. Although data oscillations result in abnormal travel patterns, their influence to the whole production pipeline should be measured. For studies with limited computation resource and relatively high data quality, this step could be ignored.

In vehicle type classification, a random forest algorithm with a support vector machine as the baseline model is implemented. Future work could be experiments with more machine learning algorithms, such as Extreme Gradient Boosting algorithm - a competitive algorithm for classification, to compare the performances. In this study,

binary classification, i.e., light-medium and heavy, is conducted for trucks. Classification by light, medium, heavy weight truck is also worth investigating. For some studies, such as those regarding urban and local delivery trucks, separating medium and light trucks may be more useful than combining them as light-medium trucks. Future work could extend this study by exploring more algorithms and more classification types. Additionally, supervised classification is feasible with the truck type information from dataset I and data segmentation (clustering) of dataset I and II. In practice, missing truck type information is a more common case from passively-collected GPS data. Future work could discuss the feasibility and performance of unsupervised classification through some other algorithms, such as deep neural network.

In the section on iterative ODME, a Q-learning algorithm is applied with the division of OD-paths into three groups, i.e., three agents. The ideal case is having as many agents as possible. Due to the extremely long computation time when setting numerous agents, this is not implemented in this present study. In the future, this might be resolved with the improvement on the algorithm or computation resources. In this study, there are three options of action for each OD-path group, i.e., 0%, 5%, 10%. This action set is roughly decided based on experience. Future work may include a discussion on the setting of action set. For example, the algorithm may perform differently regarding accuracy and efficiency with different action sets or with more options of action. As for the discussion on spatiotemporal biases of the sample GPS truck trips, there are some future work that can be conducted to enhance this research. The observed truck flows from truck count sensors on a highway network are used for

correcting the spatial bias. Hence, the representativeness of the corrected spatial distribution is limited by that of the sensors. Truck sensors are not installed everywhere on a highway network, meaning they may affect the spatial distribution bias to some extent. Future work could measure the representativeness of the truck count sensors and correct the possible spatial distribution bias of sensors before utilizing them to adjust the spatial distribution of the GPS truck flows. The temporal biases are not thoroughly discussed in the current research. Depending on the research need, temporal bias at specific resolution level may be needed. For instance, the present research conducts the monthly total estimation for January 2020. Future work may be conducted to investigate the daily representativeness of GPS data. It is possible that GPS data have inconsistent sampling rate or coverage across days, which should be measured before the deployment of the framework. In cases of deriving annual-level statistics from the selected month(s), the relation between monthly level statistics and annual level statistics should be examined. Additionally, future work can be carried out by comparing the developed framework with the existing frameworks for case studies of smaller scales. Although there is no comparable framework at a large scale as the national level in the U.S., the discussion on the accuracy and computation efficiency of the developed framework would be more complete if compared with the existing studies in the literature.

With the United States as the case study, national truck flows are obtained. This study discusses some aspects of the results from the validation and reasonableness point of view. A future study could investigate the result more thoroughly and try to reveal some interesting findings, such as the temporal and spatial patterns by truck type.

Interstate bottlenecks and congested corridors are major interests in practice. Future work could be carried out to identify these locations by projecting the estimated truck traffic flow to the highway network. Some studies attempt to convert truck traffic flow from CFS or FAF data. It would be interesting if future work could compare the estimated truck flow in this study to the estimated truck flow from CFS or FAF. There have been studies on integrating GPS technology with travel surveys. Such data, if available, have the merits of both GPS data and surveys. This current framework is developed fully based on GPS data. Future research could be conducted to explore the application of GPS-based surveys to derive truck traffic flows. Digitalization is getting widespread in transportation field. More and more trucking companies are embracing connected trucks, which enable a real-time data exchange and communication between trucks. The data produced from connected trucks provide not only timestamped locations but also speed, engine status, driver behavior, cargo status, fuel consumption, etc. As the connected trucks are getting widely applied, utilizing related data to fulfill truck travel flow estimation is a worthwhile future study.

Appendix A: Correlation Matrix

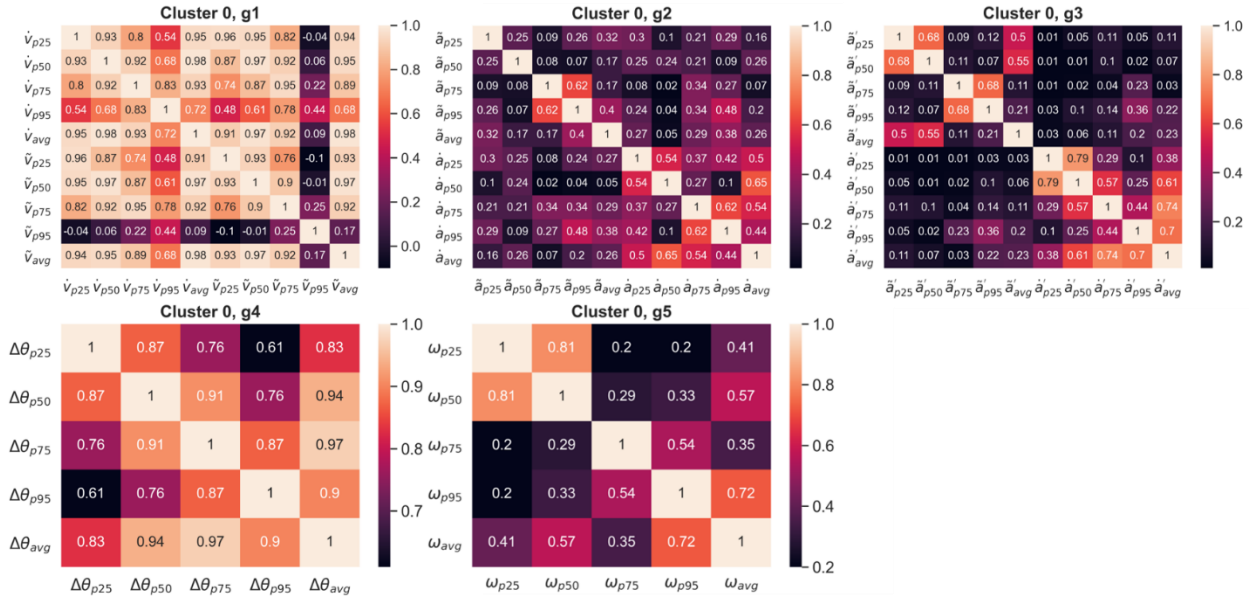


Figure 25. Correlation matrix by subgroup for cluster 0.

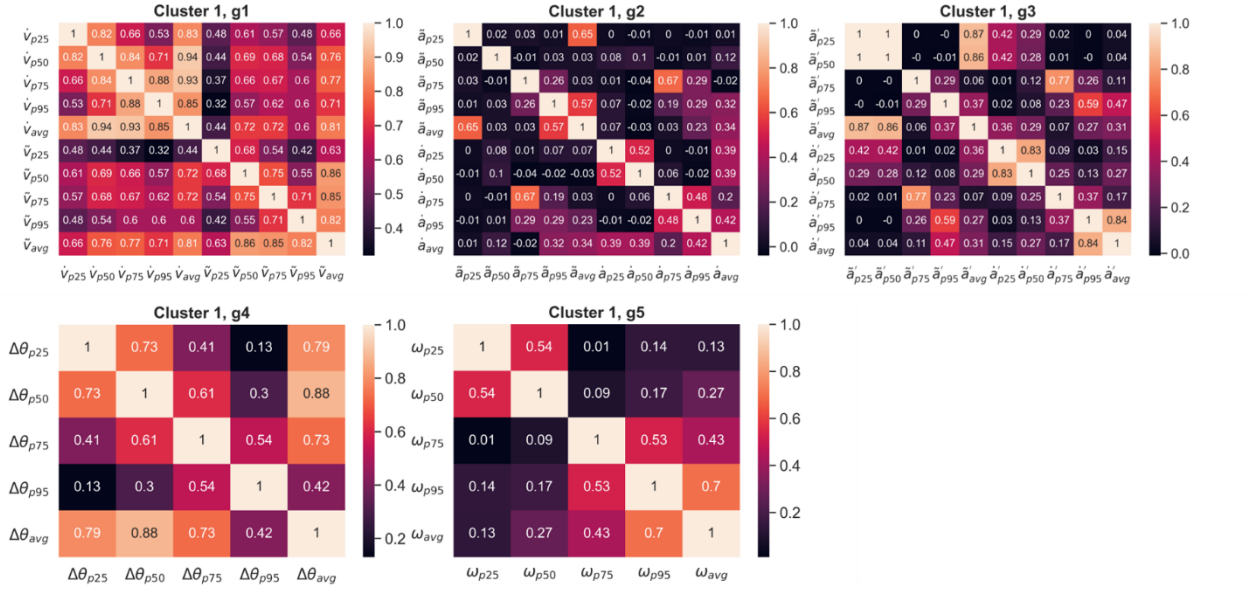


Figure 26. Correlation matrix by subgroup for cluster 1.

Appendix B: Principal Component Analysis Report

Table 12. Principal component analysis report.

Cluster 0										
g1	$\dot{v}_{p25}$	$\dot{v}_{p50}$	$\dot{v}_{p75}$	$\dot{v}_{p95}$	$\dot{v}_{avg}$	$\tilde{v}_{p25}$	$\tilde{v}_{p50}$	$\tilde{v}_{p75}$	$\tilde{v}_{p95}$	$\tilde{v}_{avg}$
pc1 (79%)	0.33	0.35	0.33	0.27	0.35	0.32	0.35	0.34	0.05	0.35
pc2 (14%)	0.21	0.07	0.15	0.44	0.04	0.27	0.15	0.14	0.78	0.02
g2	$\tilde{a}_{p25}$	$\tilde{a}_{p50}$	$\tilde{a}_{p75}$	$\tilde{a}_{p95}$	$\tilde{a}_{avg}$	$\dot{a}_{p25}$	$\dot{a}_{p50}$	$\dot{a}_{p75}$	$\dot{a}_{p95}$	$\dot{a}_{avg}$
pc1 (36%)	0.25	0.20	0.23	0.32	0.29	0.36	0.27	**0.39**	**0.39**	**0.38**
pc2 (17%)	0.07	0.20	**0.45**	**0.46**	0.20	0.28	**0.51**	0.03	0.18	0.35
pc3 (11%)	0.64	0.44	0.30	0.09	0.36	0.04	0.22	0.24	0.09	0.22
pc4 (9%)	0.07	**0.62**	**0.52**	0.19	0.25	0.07	0.21	0.14	0.41	0.08
g3	$\tilde{a}'_{p25}$	$\tilde{a}'_{p50}$	$\tilde{a}'_{p75}$	$\tilde{a}'_{p95}$	$\tilde{a}'_{avg}$	$\dot{a}'_{p25}$	$\dot{a}'_{p50}$	$\dot{a}'_{p75}$	$\dot{a}_{p95}$	$\dot{a}'_{avg}$
pc1 (32%)	0.19	0.17	0.16	0.24	0.23	0.30	**0.41**	**0.42**	0.37	**0.48**
pc2 (21%)	0.48	0.50	0.23	0.21	0.44	0.28	0.30	0.17	0.03	0.16
pc3 (16%)	0.29	0.31	0.57	0.56	0.17	0.22	0.18	0.07	0.27	0.01
pc4 (11%)	0.06	0.07	0.35	0.21	0.07	0.57	0.35	0.25	0.46	0.31
g4	$\Delta\theta_{p25}$	$\Delta\theta_{p50}$	$\Delta\theta_{p75}$	$\Delta\theta_{p95}$	$\Delta\theta_{avg}$					
pc1 (88%)	0.41	0.46	0.46	0.42	0.48					
pc2 (9%)	0.69	0.26	0.19	0.64	0.10					
g5	ω_{p25}	ω_{p50}	ω_{p75}	ω_{p95}	ω_{avg}					
pc1 (56%)	**0.43**	**0.49**	**0.36**	**0.45**	**0.50**					
pc2 (24%)	0.57	0.45	0.44	0.50	0.16					
pc3 (13%)	0.18	0.06	0.78	0.26	0.53					
Cluster 1										
g1	$\dot{v}_{p25}$	$\dot{v}_{p50}$	$\dot{v}_{p75}$	$\dot{v}_{p95}$	$\dot{v}_{avg}$	$\tilde{v}_{p25}$	$\tilde{v}_{p50}$	$\tilde{v}_{p75}$	$\tilde{v}_{p95}$	$\tilde{v}_{avg}$
pc1 (71%)	0.30	0.34	0.34	0.31	0.36	0.23	0.32	0.32	0.28	0.35
pc2 (10%)	0.10	0.23	0.33	0.35	0.28	0.65	0.32	0.20	0.15	0.20
g2	$\tilde{a}_{p25}$	$\tilde{a}_{p50}$	$\tilde{a}_{p75}$	$\tilde{a}_{p95}$	$\tilde{a}_{avg}$	$\dot{a}_{p25}$	$\dot{a}_{p50}$	$\dot{a}_{p75}$	$\dot{a}_{p95}$	$\dot{a}_{avg}$
pc1 (26%)	0.17	0.07	0.31	**0.41**	**0.39**	0.20	0.16	**0.37**	**0.41**	**0.42**
pc2 (18%)	0.06	0.16	**0.42**	0.06	0.09	**0.49**	**0.49**	0.37	0.21	0.33
pc3 (17%)	0.55	0.05	0.23	0.19	0.56	0.24	0.31	0.32	0.13	0.12
pc4 (10%)	**0.56**	0.00	**0.39**	**0.39**	0.01	0.22	0.27	0.27	0.30	0.30
g3	$\tilde{a}'_{p25}$	$\tilde{a}'_{p50}$	$\tilde{a}'_{p75}$	$\tilde{a}'_{p95}$	$\tilde{a}'_{avg}$	$\dot{a}'_{p25}$	$\dot{a}'_{p50}$	$\dot{a}'_{p75}$	$\dot{a}_{p95}$	$\dot{a}'_{avg}$
pc1 (36%)	0.42	0.41	0.14	0.21	0.46	0.33	0.33	0.18	0.25	0.26
pc2 (25%)	**0.33**	**0.33**	**0.35**	**0.37**	0.14	0.16	0.03	**0.35**	**0.46**	**0.37**
pc3 (15%)	0.08	0.08	0.45	0.26	0.24	0.33	0.39	0.47	0.28	0.32
pc4 (13%)	0.25	0.25	0.41	0.04	0.20	0.47	0.50	0.30	0.10	0.30
g4	$\Delta\theta_{p25}$	$\Delta\theta_{p50}$	$\Delta\theta_{p75}$	$\Delta\theta_{p95}$	$\Delta\theta_{avg}$					

pc1 (66%)	0.44	0.50	0.44	0.29	0.53					
pc2 (20%)	0.48	0.23	0.34	0.77	0.11					
g5	ω_{p25}	ω_{p50}	ω_{p75}	ω_{p95}	ω_{avg}					
pc1 (46%)	0.26	0.33	0.45	0.56	0.55					
pc2 (28%)	0.66	0.60	0.34	0.25	0.16					
pc3 (12%)	0.14	0.07	0.81	0.24	0.52					

Bibliography

[1] Abdull, N., Yoneda, M., & Shimada, Y. (2020). Traffic characteristics and pollutant emission from road transport in urban area. *Air Quality, Atmosphere & Health*, *13*(6), 731–738. https://doi.org/10.1007/s11869-020-00830-w

[2] Advanced Solutions International, Inc. (2020). *31st Annual State of Logistics Report*. https://cscmp.org/store/detail.aspx?id=SOL-20

[3] Advanced Solutions International, Inc. (2021). *32nd Annual State of Logistics Report*. https://cscmp.org/store/detail.aspx?id=SOL-21

[4] Advanced Solutions International, Inc. (2022). *CSCMP's 33rd Annual State of Logistics Report*. https://cscmp.org/store/detail.aspx?id=SOL-22

[5] Allen, J., Ambrosini, C., Browne, M., Patier, D., Routhier, J., & Woodburn, A. G. (2014). Data Collection for Understanding Urban Goods Movement. In *Ecoproduction* (pp. 71–89). Springer International Publishing. https://doi.org/10.1007/978-3-642-31788-0_5

[6] Aziz, R., Kedia, M., Dan, S., Basu, S., Sarkar, S., Mitra, S., & Mitra, P. (2016). Identifying and Characterizing Truck Stops from GPS Data. In *Lecture Notes in Computer Science* (pp. 168–182). Springer Science+Business Media. https://doi.org/10.1007/978-3-319-41561-1_13

[7] Barnett, V., & Lewis, T. (1994). *Outliers in Statistical Data*. John Wiley & Sons.

[8] Bernardin, V. L., Avner, J., Short, J., Brown, L., Nunnally, R., & Smith, S. (2011). Using large sample GPS data to develop an improved truck trip table for the Indiana statewide model. *TRB Innovation Papers*.

[9] Bernardin, V. L., Jr, Ferdous, N., Sadrsadat, H., Trevino, S., & Chen, C. (2017). Integration of National Long-Distance Passenger Travel Demand Model with Tennessee Statewide Model and Calibration to Big Data. *Transportation Research Record*, *2653*(1), 75–81. https://doi.org/10.3141/2653-09

[10] Breunig, M., Kriegel, H., Ng, R. T., & Sander, J. (2000). LOF. *Sigmod Record*, *29*(2), 93–104. https://doi.org/10.1145/335191.335388

[11] Bricka, S., Sen, S., Paleti, R., & Bhat, C. R. (2012). An analysis of the factors influencing differences in survey-reported and GPS-recorded trips. *Transportation Research Part C-emerging Technologies*, *21*(1), 67–88. https://doi.org/10.1016/j.trc.2011.09.005

[12] Calabrese, F., Di Lorenzo, G., Liu, L., & Ratti, C. (2011). Estimating Origin-Destination Flows Using Mobile Phone Location Data. *IEEE Pervasive Computing*, *10*(4), 36–44. https://doi.org/10.1109/mprv.2011.41

[13] Camargo, P., Hong, S., & Livshits, V. (2017). Expanding the Uses of Truck GPS Data in Freight Modeling and Planning Activities. *Transportation Research Record*, *2646*(1), 68–76. https://doi.org/10.3141/2646-08

[14] Chen, C., Bian, L., & Jingtao, M. (2014). From traces to trajectories: How well can we guess activity locations from mobile phone traces? *Transportation Research*

Part C-emerging Technologies, *46*, 326–337. https://doi.org/10.1016/j.trc.2014.07.001

[15] Chen, C., Zhao, X., Zhang, Y., Rong, J., & Liu, X. (2019). A graphical modeling method for individual driving behavior and its application in driving safety analysis using GPS data. *Transportation Research Part F-traffic Psychology and Behaviour*, *63*, 118–134. https://doi.org/10.1016/j.trf.2019.03.017

[16] Chen, W., Ji, M., & Wang, J. (2014). T-DBSCAN: A Spatiotemporal Density Clustering for GPS Trajectory Segmentation. *International Journal of Online and Biomedical Engineering*, *10*(6), 19. https://doi.org/10.3991/ijoe.v10i6.3881

[17] Comi, A., Coppola, P., & Nuzzolo, A. (2013). Freight transport modeling: review and future challenges. *Freight Transport Modeling: Review and Future Challenges*, 151-182.

[18] Coifman, B., & Kim, S. (2009). Speed estimation and length based vehicle classification from freeway single-loop detectors. *Transportation Research Part C-emerging Technologies*, *17*(4), 349–364. https://doi.org/10.1016/j.trc.2009.01.004

[19] Dabiri, S., Markovic, N. M., Heaslip, K., & Reddy, C. K. (2020). A deep convolutional neural network based approach for vehicle classification using large-scale GPS trajectory data. *Transportation Research Part C-emerging Technologies*, *116*, 102644. https://doi.org/10.1016/j.trc.2020.102644

[20] Daniel, J., Tsai, C., & Chien, S. (2002). Factors in truck crashes on roadways with intersections. *Transportation research record*, *1818*(1), 54-59.

[21] Demissie, M. G., & Kattan, L. (2022). Estimation of truck origin-destination flows using GPS data. *Transportation Research Part E-logistics and Transportation Review*, *159*, 102621. https://doi.org/10.1016/j.tre.2022.102621

[22] Dong, W., Li, J., Yao, R., Li, C., Yuan, T., & Wang, L. (2016). Characterizing driving styles with deep learning. arXiv 2016. *arXiv preprint arXiv:1607.03611*.

[23] Eikvil, L., Aurdal, L., & Koren, H. (2009). Classification-based vehicle detection in high-resolution satellite images. *Isprs Journal of Photogrammetry and Remote Sensing*, *64*(1), 65–72. https://doi.org/10.1016/j.isprsjprs.2008.09.005

[24] Ester, M., Kriegel, H., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial Databases with Noise. In *Knowledge Discovery and Data Mining* (pp. 226–231). https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf

[25] Ferreira, F., Wojtak, W., Fernandes, C. A., Guimarães, P. M. R., Monteiro, S. N., Bicho, E., & Erlhagen, W. (2021). Dynamic Identification of Stop Locations from GPS Trajectories Based on Their Temporal and Spatial Characteristics. In *Lecture Notes in Computer Science*. Springer Science+Business Media. https://doi.org/10.1007/978-3-030-86380-7_28

[26] *FHWA Operations - Office of Freight Management and Operations - Freight Analysis Framwork (FAF) Plan*. (n.d.). https://ops.fhwa.dot.gov/docs/fafplandraft/fafplandraft.htm

[27] Federal Highway Administration. (2022). *2022 FHWA Forecasts of Vehicle Miles Traveled (VMT)* [White paper]. Federal Highway Administration. https://www.fhwa.dot.gov/policyinformation/tables/vmt/2022_vmt_forecast_sum.pdf

[28] *Highway Statistics Series - Policy | Federal Highway Administration*. (n.d.). https://www.fhwa.dot.gov/policyinformation/statistics.cfm

[29] Fischer, M., Outwater, M. L., Cheng, L. L., Ahanotu, D. N., & Calix, R. (2005). Innovative Framework for Modeling Freight Transportation in Los Angeles County, California. *Transportation Research Record*, *1906*(1), 105–112. https://doi.org/10.3141/1906-13

[30] Freire, M. R., Gauld, C., McKerral, A., & Pammer, K. (2021). Identifying Interactive Factors That May Increase Crash Risk between Young Drivers and Trucks: A Narrative Review. *International Journal of Environmental Research and Public Health*, *18*(12), 6506. https://doi.org/10.3390/ijerph18126506

[31] Fu, Z., Tian, Z., Xu, Y., & Qiao, C. (2016). A Two-Step Clustering Approach to Extract Locations from Individual GPS Trajectory Data. *ISPRS International Journal of Geo-information*, *5*(10), 166. https://doi.org/10.3390/ijgi5100166

[32] Gingerich, K., Maoh, H., & Anderson, W. F. (2016). Classifying the purpose of stopped truck events: An application of entropy to GPS data. *Transportation Research Part C-emerging Technologies*, *64*, 17–27. https://doi.org/10.1016/j.trc.2016.01.002

[33] Gong, L., Sato, H., Yamamoto, T., Miwa, T., & Morikawa, T. (2015). Identification of activity stop locations in GPS trajectories by density-based clustering method combined with support vector machines. *Journal of Modern Transportation*, *23*(3), 202–213. https://doi.org/10.1007/s40534-015-0079-x

[34] Gong, L., Yamamoto, T., & Morikawa, T. (2018). Identification of activity stop locations in GPS trajectories by DBSCAN-TE method combined with support vector machines. *Transportation Research Procedia*, *32*, 146–154. https://doi.org/10.1016/j.trpro.2018.10.028

[35] Guha, S., Rastogi, R., & Shim, K. (2001). Cure: an efficient clustering algorithm for large databases. *Information Systems*, *26*(1), 35–58. https://doi.org/10.1016/s0306-4379(01)00008-4

[36] Guo, J., Liu, Y., Zhang, L., & Wang, Y. (2018). Driving Behaviour Style Study with a Hybrid Deep Learning Framework Based on GPS Data. *Sustainability*, *10*(7), 2351. https://doi.org/10.3390/su10072351

[37] Gupta, M., Gao, J., Aggarwal, C. C., & Han, J. (2014). Outlier Detection for Temporal Data. *Synthesis Lectures on Data Mining and Knowledge Discovery*, *5*(1), 1–129. https://doi.org/10.2200/s00573ed1v01y201403dmk008

[38] Harwood, D. W., Torbic, D. J., Richard, K. R., Glauz, W. D., & Elefteriadou, L. (2003). Review of Truck Characteristics as Factors in Roadway Design. In *Transportation Research Board eBooks*. https://doi.org/10.17226/23379

[39] Hawkins, D. M. (1980). Identification of Outliers. In *Springer eBooks*. https://doi.org/10.1007/978-94-015-3994-4

[40] Horn, C. B., Klampfl, S., Cik, M., & Reiter, T. (2014). Detecting Outliers in Cell Phone Data. *Transportation Research Record*, *2405*(1), 49–56. https://doi.org/10.3141/2405-07

[41] Hossan, S., Asgari, H., & Jin, X. (2018). Trip misreporting forecast using count data model in a GPS enhanced travel survey. *Transportation*, *45*(6), 1687–1700. https://doi.org/10.1007/s11116-017-9782-2

[42] IBI Group. (2009). *Final Report: Continental Gateway Road Network Performance*. Transport Canada, February.

[43] Iovan, C., Olteanu-Raimond, A., Couronné, T., & Smoreda, Z. (2013). Moving and Calling: Mobile Phone Data Quality Measurements and Spatiotemporal Uncertainty in Human Mobility Studies. In *Lecture notes in geoinformation and cartography* (pp. 247–265). Springer Nature. https://doi.org/10.1007/978-3-319-00615-4_14

[44] Jacob, O. O., Chukwudi, I. C., Thaddeus, E. O., & Agwu, E. C. (2020). Estimation of the Impact of the Overloaded Truck on the Service Life of Pavement Structures in Nigeria. *International Journal for Traffic and Transport Engineering*, *9*(2), 41–47. http://article.sapub.org/10.5923.j.ijtte.20200902.03.html

[45] Jolliffe, I. (2013). *Principal Component Analysis*. Springer Science & Business Media.

[46] Kafai, M., & Bhanu, B. (2012). Dynamic Bayesian Networks for Vehicle Classification in Video. *IEEE Transactions on Industrial Informatics*, *8*(1), 100–109. https://doi.org/10.1109/tii.2011.2173203

[47] Knorr, E. M., Ng, R. T., & Tucakov, V. (2000). Distance-based outliers: algorithms and applications. *The Vldb Journal*, *8*(3–4), 237–253. https://doi.org/10.1007/s007780050006

[48] Kong, D., Guo, X., Yang, B., & Wu, D. (2016). Analyzing the Impact of Trucks on Traffic Flow Based on an Improved Cellular Automaton Model. *Discrete Dynamics in Nature and Society*, *2016*, 1–14. https://doi.org/10.1155/2016/1236846

[49] Krishnan, S., Garg, A., Patil, S., Lea, C., Hager, G., Abbeel, P., & Goldberg, K. (2017). Transition state clustering: Unsupervised surgical trajectory segmentation for robot learning. *The International Journal of Robotics Research*, *36*(13-14), 1595-1618.

[50] Kuppam, A. R., R, C., Lemp, J., Rossi, T., Livshits, V., Vallabhaneni, L., Jeon, K., & Brown, E. R. (2013). Special events travel surveys and model development. *Transportation Letters: The International Journal of Transportation Research*, *5*(2), 67–82. https://doi.org/10.1179/1942786713z.0000000007

[51] Kuppam, A., Lemp, J., Beagan, D., Livshits, V., Nippani, S., & Vallabhaneni, L. (2014). Development of a tour-based truck travel demand model. *Innov. Travel model*.

[52] Laranjeiro, P. F., Pardalos, P. M., Godoy, L., Giannotti, M. A., Yoshizaki, H. T. Y., Winkenbach, M., & Da Cunha, C. B. (2019). Using GPS data to explore speed patterns and temporal fluctuations in urban logistics: The case of São Paulo, Brazil. *Journal of Transport Geography*, *76*, 114–129. https://doi.org/10.1016/j.jtrangeo.2019.03.003

[53] Lee, J., & Hou, J. C. (2006). *Modeling steady-state and transient behaviors of user mobility*. https://doi.org/10.1145/1132905.1132915

[54] Lee, J. G., Han, J., & Li, X. (2008, April). Trajectory outlier detection: A partition-and-detect framework. In *2008 IEEE 24th International Conference on Data Engineering* (pp. 140-149). IEEE.

[55] Lee, M. H., Zhao, J., Sun, Q., Pan, Y., Zhou, W., Xiong, C., & Zhang, L. (2020). Human mobility trends during the early stage of the COVID-19 pandemic in the United States. *PLOS ONE*, *15*(11), e0241468. https://doi.org/10.1371/journal.pone.0241468

[56] Li, X., Han, J., Kim, S., & Gonzalez, H. (2007, April). Roam: Rule-and motif-based anomaly detection in massive moving object data sets. In *Proceedings of the 2007 SIAM International Conference on Data Mining* (pp. 273-284). Society for Industrial and Applied Mathematics.

[57] Lopatin, O. P. (2020). Development of a methodology for assessing the impact of vehicles on the acoustic environment. *IOP Conference Series*, *548*(6), 062049. https://doi.org/10.1088/1755-1315/548/6/062049

[58] Ma, X., & Grimson, W. E. L. (2005). *Edge-based rich representation for vehicle classification*. https://doi.org/10.1109/iccv.2005.80

[59] Ma, X., McCormack, E., & Wang, Y. (2011). Processing Commercial Global Positioning System Data to Develop a Web-Based Truck Performance Measures Program. *Transportation Research Record*, *2246*(1), 92–100. https://doi.org/10.3141/2246-12

[60] Ma, Y., Van Zuylen, H. J., & Kuik, R. (2012). *Freight origin-destination estimation based on multiple data source*. https://doi.org/10.1109/itsc.2012.6338625

[61] McGowen, P. T. (2006). *Predicting activity types from GPS and GIS data*. University of California, Irvine.

[62] Montini, L., Rieser-Schüssler, N., Horni, A., & Axhausen, K. W. (2014). Trip Purpose Identification from GPS Tracks. *Transportation Research Record*, *2405*(1), 16–23. https://doi.org/10.3141/2405-03

[63] *National Transportation Statistics*. (2021). Bureau of Transportation Statistics. https://www.bts.gov/topics/national-transportation-statistics

[64] Ni, L., Wang, X., & Chen, X. (2018). A spatial econometric model for travel flow analysis and real-world applications with massive mobile phone data. *Transportation Research Part C-emerging Technologies*, *86*, 510–526. https://doi.org/10.1016/j.trc.2017.12.002

[65] Palma, A. T., Bogorny, V., Kuijpers, B., & Alvares, L. O. (2008). *A clustering-based approach for discovering interesting places in trajectories*. https://doi.org/10.1145/1363686.1363886

[66] Papadimitriou, S., Kitagawa, H., Gibbons, P. B., & Faloutsos, C. (2003). *LOCI: fast outlier detection using the local correlation integral*. https://doi.org/10.1109/icde.2003.1260802

[67] Ross, Z., Kheirbek, I., Clougherty, J. E., Ito, K., Matte, T., Markowitz, S. M., & Eisl, H. (2011). Noise, air pollutants and traffic: Continuous measurement and correlation at a high-traffic location in New York City. *Environmental Research*, *111*(8), 1054–1063. https://doi.org/10.1016/j.envres.2011.09.004

[68] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

[69] Sarti, L., Bravi, L., Sambo, F., Taccari, L., Simoncini, M., Salti, S., & Lori, A. (2017). *Stop Purpose Classification from GPS Data of Commercial Vehicle Fleets*. https://doi.org/10.1109/icdmw.2017.43

[70] Sarvi, M. (2011). Heavy commercial vehicles-following behavior and interactions with different vehicle classes. *Journal of Advanced Transportation*, *47*(6), 572–580. https://doi.org/10.1002/atr.182

[71] Sharpe, B. E. N., & Muncrief, R. (2015). Literature review: real-world fuel consumption of heavy-duty vehicles in the United States, China, and the European Union. *International Council on Clean Transportation (ICCT), Washington DC.*

[72] Simoncini, M., Sambo, F., Taccari, L., Bravi, L., Salti, S., & Lori, A. (2016). *Vehicle Classification from Low Frequency GPS Data*. https://doi.org/10.1109/icdmw.2016.0167

[73] Simoncini, M., Taccari, L., Sambo, F., Bravi, L., Salti, S., & Lori, A. (2018). Vehicle classification from low-frequency GPS data with recurrent neural networks. *Transportation Research Part C-emerging Technologies*, *91*, 176–191. https://doi.org/10.1016/j.trc.2018.03.024

[74] Sun, D., Leurent, F., & Xie, X. (2020). Floating Car Data mining: Identifying vehicle types on the basis of daily usage patterns. *Transportation Research Procedia*, *47*, 147–154. https://doi.org/10.1016/j.trpro.2020.03.087

[75] Sun, Q., Zhou, W., Kabiri, A., Darzi, A., Hu, S., Younes, H., & Zhang, L. (2022). COVID-19 and income profile: How communities in the United States responded to mobility restrictions in the pandemic's early stages. *Regional Science Policy and Practice*, *15*(3), 541–558. https://doi.org/10.1111/rsp3.12598

[76] Sun, Z., & Ban, X. (2013). Vehicle classification using GPS data. *Transportation Research Part C-emerging Technologies*, *37*, 102–117. https://doi.org/10.1016/j.trc.2013.09.015

[77] Taylor, B. D., Miller, D. C., Iseki, H., & Fink, C. (2009). Nature and/or nurture? Analyzing the determinants of transit ridership across US urbanized areas. *Transportation Research Part A-policy and Practice*, *43*(1), 60–77. https://doi.org/10.1016/j.tra.2008.06.007

[78] Thakur, A., Pinjari, A. R., Zanjani, A. B., Short, J. W., Mysore, V., & Tabatabaee, S. F. (2015). Development of Algorithms to Convert Large Streams of Truck GPS Data into Truck Trips. *Transportation Research Record*, *2529*(1), 66–73. https://doi.org/10.3141/2529-07

[79] Thiagarajan, A., Ravindranath, L., LaCurts, K., Madden, S., Balakrishnan, H., Toledo, S., & Eriksson, J. (2009, November). Vtrack: accurate, energy-aware road traffic delay estimation using mobile phones. In *Proceedings of the 7th ACM conference on embedded networked sensor systems* (pp. 85-98).

[80] Tran, L. H., Nguyen, Q. V. H., Do, N. H., & Yan, Z. (2011). *Robust and hierarchical stop discovery in sparse and diverse trajectories* (No. EPFL-REPORT-175473).

[81] Tseng, Y. Y., Yue, W. L., & Taylor, M. A. (2005, June). The role of transportation in logistics chain. Eastern Asia Society for Transportation Studies.

[82] U.S. Department of Transportation. (2022, February). *Supply Chain Assessment of the Transportation Industrial Base: Freight and Logistic*s [White paper]. U.S. Department of Transportation. https://www.transportation.gov/sites/dot.gov/files/2022-

02/EO%2014017%20-%20DOT%20Sectoral%20Supply%20Chain%20Assessment%20-%20Freight%20and%20Logistics_FINAL.pdf

[83] Vital, F., Ioannou, P., & Gupta, A. (2021). Survey on Intelligent Truck Parking: Issues and Approaches. *IEEE Intelligent Transportation Systems Magazine*, *13*(4), 31–44. https://doi.org/10.1109/mits.2019.2926259

[84] Wang, F., & Chen, C. (2018). On data processing required to derive mobility patterns from passively-generated mobile phone data. *Transportation Research Part C-emerging Technologies*, *87*, 58–74. https://doi.org/10.1016/j.trc.2017.12.003

[85] Wu, W., Wang, Y., Gomes, J., Antonatos, S., Xue, M., Yang, P., Yap, G., Li, X., Krishnaswamy, S., Decraene, J., & Nash, A. S. (2014). *Oscillation Resolution for Mobile Phone Cellular Tower Data to Enable Mobility Modelling*. https://doi.org/10.1109/mdm.2014.46

[86] Yang, F., Wang, Y., Jin, P. J., Li, D., & Yao, Z. (2021). Random Forest Model for Trip End Identification Using Cellular Phone and Points of Interest Data. *Transportation Research Record*, *2675*(7), 454–466. https://doi.org/10.1177/03611981211031537

[87] Yang, X., & Tang, L. (2016). CROWDSOURCING BIG TRACE DATA FILTERING: A PARTITION-AND-FILTER MODEL. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, *41*.

[88] You, S. I., & Ritchie, S. G. (2019). Tour-Based Truck Demand Modeling with Entropy Maximization Using GPS Data. *Journal of Advanced Transportation*, *2019*, 1–11. https://doi.org/10.1155/2019/5021026

[89] Yu, Y., Cao, L., Rundensteiner, E. A., & Wang, A. (2014). *Detecting moving object outliers in massive-scale trajectory streams*. https://doi.org/10.1145/2623330.2623735

[90] Yuksel, E., Bertini, R. L., Menon, N., Ozkul, S., & Staes, B. (2020). *A Contemporary Approach for Visualizing Temporal and Spatial Urban Freight Movement by Leveraging Mobility Portal Data*. https://doi.org/10.1109/fists46898.2020.9264848

[91] Zanjani, A. B., Pinjari, A. R., Capela, I., Thakur, A., Short, J. W., Mysore, V., & Tabatabaee, S. F. (2015). Estimation of Statewide Origin–Destination Truck Flows from Large Streams of GPS Data. *Transportation Research Record*, *2494*(1), 87–96. https://doi.org/10.3141/2494-10

[92] Zhang, L., Darzi, A., Ghader, S., Pack, M. L., Xiong, C., Yang, M., Sun, Q., Kabiri, A., & Hu, S. (2021a). Interactive COVID-19 Mobility Impact and Social Distancing Analysis Platform. *Transportation Research Record*, 036119812110438. https://doi.org/10.1177/03611981211043813

[93] Zhang, L., & Wang, Z. (2011). Trajectory Partition Method with Time-Reference and Velocity. *JCIT, AICIT*, *6*(8), 134-142.