Understanding Freight Trip-Chaining Behavior Using a Spatial Data-Mining Approach with GPS Data

Xiaolei Ma, Yong Wang, Edward McCormack, and Yinhai Wang

Freight systems are a critical yet complex component of the transportation domain. Understanding the dynamic of freight movements will help in better management of freight demand and eventually improve freight system efficiency. This paper presents a series of data-mining algorithms to extract an individual truck's trip-chaining information from multiday GPS data. Individual trucks' anchor points were identified with the spatial clustering algorithm for density-based spatial clustering of applications with noise. The anchor points were linked to construct individual trucks' trip chains with 3-day GPS data, which showed that 51% of the trucks in the data set had at least one trip chain. A partitioning around medoids nonhierarchical clustering algorithm was applied to group trucks with similar trip-chaining characteristics. Four clusters were generated and validated by visual inspection when the trip-chaining statistics were distinct from each other. This study sheds light on modeling freight-chaining behavior in the context of massive freight GPS data sets. The proposed trip chain extraction and behavior classification algorithms can be readily implemented by transportation researchers and practitioners to facilitate the development of activity-based freight demand models.

Freight transportation plays a critical role in the U.S. economy. Trucks carry the largest market share of all freight modes (*1*). In the past few decades, as freight activity has grown and the role of the trucking industry has become more critical, the need to manage freight demand, alleviate roadway congestion, and monitor system performance has become increasingly urgent (*2*). In 2012, FHWA began implementation of the Moving Ahead for Progress in the 21st Century Act (MAP-21) to improve the efficiency, productivity, and resilience of the American transportation system. MAP-21 aims to develop a multimodal and performance-based program to address systematically the multiple critical needs of the U.S. transportation system (*3*). For the freight sector, MAP-21 emphasizes establishing a national freight network to direct limited resources for improving freight mobility on highways (*3*). As a major component of MAP-21, a series of strategic freight planning tools are recommended, including demand forecasting, performance measurements, and bottleneck identification. Without an in-depth understanding of freight travel patterns and the proper modeling of freight activities, the development of effective and comprehensive freight planning tools is challenging.

Activity-based freight models can support freight planning and demand management, but there have been few applications. In the past two decades, several activity-based models have been proposed and developed for passenger trip demand forecasting. Recently, these models have seen increased attention by transportation researchers and practitioners (*4, 5*). The activity-based approaches can forecast travel demand derived from individual activities whose patterns are typically considered as the basic unit of analysis (*6–10*). Compared with the conventional four-step trip-based model, activity-based models reflect the links between trips and activities and can potentially better estimate urban commercial vehicle flows (*11*). Compared with passenger travel, freight activities are more complex to model, because of the reasons outlined by Holguín-Veras and Patil (*12*).

An important difference between freight and passenger travel models is the need to account for the dependency between vehicle trips. Trucks often visit multiple customers during their deliveries (*12*), while passenger auto trips may typically include only one or two stops (*13*). The assumption behind conventional demand forecasting models is that trips made in one long tour are considered independent of each other (*14*), but this is not reasonable for freight models. The interaction between different truck trips reflects the actual demand as created by multiple decision makers. For this reason, the trip-chaining behaviors that are inherent in freight mobility should be taken into account.

Data scarcity is an ongoing limitation on the development of freight performance measures and activity-based freight demand models (*15*). Freight activity information (e.g., trip tour and travel patterns) need to be used to model freight movements accurately (*16*). Freight activity information is often difficult to acquire because of concerns related to the protection of customers' privacy and business competitiveness. Traditionally, most freight demand forecasting studies have relied on manual data collection methods, such as travel diaries or logs and surveys. These methods are fairly costly and difficult to implement at a multiday level, because of low response rates and inaccuracy (*17*).

A growing number of trucking companies have resorted to GPS technology for fleet management (*18*). It has been estimated that the number of fleet management systems in active use in North America has grown from 2.8 million units in 2011 to 5.9 million units by 2016 (*19*). Trucking companies' GPS data are mainly designed for monitoring truck operations and driver behaviors rather than for estimating freight demand (*20*). The data have quality problems,

X. Ma, Department of Transportation Engineering, School of Transportation Science and Engineering, Beijing Key Laboratory for Cooperative Vehicle Infrastructure, Systems, and Safety Control, Beihang University, Beijing 100191, China. Yong Wang, School of Management, Chongqing Jiaotong University, Chongqing 400074, China. E. McCormack and Yinhai Wang, Department of Civil and Environmental Engineering, College of Engineering, University of Washington, Box 352700, Seattle, WA 98195-2700. Corresponding author: E. McCormack, edm@uw.edu.

Transportation Research Record: Journal of the Transportation Research Board, No. 2596, Transportation Research Board, Washington, D.C., 2016, pp. 44–54.
DOI: 10.3141/2596-06

such as GPS signal loss and fluctuation, and lack information related to freight activity, such as trip origin–destination (O-D) and trip chains. Analytical approaches that eliminate erroneous data, adjust for missing values, and discover hidden and indirect information are necessary to improve the usability of truck GPS data.

The methodology presented in this paper uses a series of data-mining algorithms to extract trip-chaining information from a large set of truck GPS data and then uses these data to model truckers' chaining behaviors. A spatial clustering algorithm is initially applied to identify clusters of geospatial points where truck depots are likely located. Then, trip-chaining information (e.g., number of trip chains, dwell time, and number of trips) is calculated for individual trucks on a daily basis. From the extracted trip-chaining features, a nonhierarchical clustering algorithm is used to categorize truck travel patterns into different groups with distinct characteristics. The contributions of this study are twofold: (*a*) providing an efficient and effective spatial data-mining approach to extract freight activity information (e.g., trip O-D and trip chain) from massive GPS data sets; and (*b*) proposing a partitioning around medoids (PAM) clustering algorithm to categorize truck trip-chaining behaviors.

This paper is organized as follows. The relevant research on freight trip-chaining modeling is initially summarized, followed by a brief introduction of the GPS data set and geospatial network used for the study. The paper describes the development of several data-mining approaches for trip chain identification and freight-chaining behavior classification. Three-day GPS data were used to test the effectiveness of the proposed algorithms. Finally, the limitations of the methodology are discussed, and future research directions are suggested.

LITERATURE REVIEW

Various disaggregate activity-based models have been developed to capture freight activities (*21*). In most of these models, a trip tour (or chain) is defined as a sequence of trips. These trips start from a base depot or warehouse and end at the same location after several deliveries to customers (*4, 22*). Tour-based freight models can more accurately describe the dependency between freight trips and therefore are better at explaining the dynamics of freight activities. Holguín-Veras and Patil (*12*) analyzed the trip-chaining behavior of commercial vehicles with data from Denver, Colorado. In their study, the number of trip chains, trip chain length, and the probability of a specific trip purpose given the relative stop location were used to characterize trip-chaining behavior. The study found that most trucks have at least one trip chain per day, and the number of stops per trip chain decreases with the number of trip chains.

Hunt and Stefan (*23*) developed a tour-based microsimulation system to model urban commercial movements for Calgary, Alberta, Canada. Wang and Holguín-Veras (*14*) proposed two models that used data from a freight microsimulation tool to determine trip tour destinations and the time when a trip tour is terminated. Several factors were found to influence trip-chaining behavior, including the choice of the next destination and the distance from the current location to the potential destination. Figliozzi (*16*) analyzed commercial vehicle tours by changing various routing constraints and found that the distribution of trip length largely relies on the tour type, distance from the depot, and number of stops per tour. In addition, the author pointed out that automatic freight behavioral data collection is necessary for freight demand modeling. Figliozzi (*24*) later used empirical tour data from Sydney, Australia, to classify trip tours into three groups based on tour efficiency, time, and distance-related costs.

These trip tour research efforts relied heavily on manual data collection methods (e.g., survey and travel diary) and simulation tools for analyzing freight-chaining behavior. GPS-equipped devices have become more ubiquitous, and the freight industry has taken advantage of this technology to track and manage trucking operations. The use of passive and secondary freight data from these devices provides another opportunity to improve freight behavioral studies.

GPS data can be utilized by transportation researchers to understand passenger travel behavior in the past. For example, Li et al. (*25*) used 1-week GPS data to investigate morning commuters' travel patterns. Du and Aultman-Hall (*26*) developed a heuristic model with heading change, dwell, and distance between GPS points to detect trip ends based on multiday GPS travel data sets, and the algorithm accurately captured 94% of trip ends. Alvarez-Garcia et al. (*27*) developed a Hidden Markov Model to predict individual travelers' trip destinations based on GPS data. Spissu et al. (*28*) analyzed the route choices of 12 respondents from GPS data collected over 2 weeks. Greaves and Figliozzi (*20*) discussed the issues and potential applications of GPS data for commercial vehicle tour information extraction. The authors proposed a method to process raw GPS data from Melbourne, Australia, into trip information (e.g., trip destination and trip tour). Sharman and Roorda (*17*) applied a hierarchical agglomeration and partitioning clustering algorithms to identify GPS trip ends in repeated destinations.

These studies show the promise of utilizing GPS data to collect freight activity information. However, little research has been completed with the use of GPS data to develop an effective and efficient approach to extract individual truck trip tour information and classify trucks' activities with their different trip-chaining behaviors. This study attempted to address these issues by introducing several data-mining algorithms.

DATA SET

The truck data set used in this study was obtained under a contract with GPS fleet management device vendors, with each vendor providing GPS data from multiple trucking companies (*18*). The vendors were willing to share their GPS data with a university, since the resulting freight performance measures and potential infrastructure improvements could be beneficial for them.

Starting in 2008, GPS data from several trucking companies traveling in Washington State were fed into a server at the University of Washington. Because of the increasing size of GPS data over time, several relational databases were designed to archive these data sets. An automatic data retrieval program was developed to fetch the raw files from vendors' FTP servers on a daily basis, and the data were parsed and imported into databases. On average, there are 3,400 trucks every weekday. Each probe truck, when moving, transmits a GPS signal with a cellular connection approximately every 5 min (this is not applicable when a truck idles or parks).

The GPS information includes the device identifier, engine status, speed, latitude, longitude, GPS status, timestamp, and kilometers traveled. The GPS device identifier was scrambled (hashed) for privacy protection. The engine status variable defines whether each vehicle's engine is on or off. The GPS status provides the strength of the GPS connection for each device, since the signal may vary or be lost when a truck travels in urban areas where high buildings degrade satellite communications. More detailed information on the data can be found in McCormack and Zhao (*29*). Data from five typical weekdays (from May 6 to May 10, 2013) were collected for this study. The data set contained more than two million GPS records.

In addition to the passively collected GPS data sets, geospatial data from Washington State were also incorporated into the study. Digital freeway data can be overlaid with the freight GPS data to determine when the trucks travel on freeways. The locations of rest areas and weigh stations were used to filter out many of the truckers' non-cargo-related stops. Kilometers recorded in GPS data can be utilized to calculate the total distance each truck travels. All of the geospatial data were stored in PostgreSQL, which is an open-source spatial database management tool.

METHODOLOGY

The definition of a freight trip chain (tour) is presented in Figure 1. A vehicle leaves from an anchor point (e.g., depot) base depot at the beginning of the workday, visits a sequence of stops, and finally returns to the anchor point at the end of the day. The entire closed-loop stop sequence is a trip chain, and a segment connecting two adjacent stops is a trip. As shown in Figure 1, the trip chain can be defined as Base→1→2→3→4→5→Base. A trip chain comprises several trips and there could be multiple daily trip chains (closed-loop stop sequences) for each vehicle (*12*).

A three-step data processing procedure was developed to extract the trip-chaining information from the raw GPS data. The anchor points (the base depots or origins) for each of the trucks can be inferred with a spatial data-mining algorithm. Finally, individual trip chains can be aggregated from the identified truck trips.

Trip O-D Estimation

An approach developed by Ma et al. (*18*) was used to estimate individual trucks' trip ends with the objective of differentiating non-delivery-related stops caused by traffic congestion and intended stops for fulfilling certain tasks (e.g., delivery and pickup services). The trip end identification approach is composed of several algorithms. The first algorithm uses a dwell time threshold (3 min) to filter out truck stops for traffic signals. This is because it is unusual for a truck to have a full stop for more than 3 min because of traffic conditions in the Puget Sound region (*30*). The second algorithm uses engine status information from the raw GPS data set to determine each truck's movements. Any stops with the engine off are considered an origin or destination. The third algorithm eliminates abnormal trips, such as extremely high travel speed and travel times, with a zero value. For individual trucks, the inferred origins and destinations are linked together, and the truck's trip can be identified as well. Of the two million records in the GPS data, 250,484 records (12.4%) were identified as trip origins or destinations.

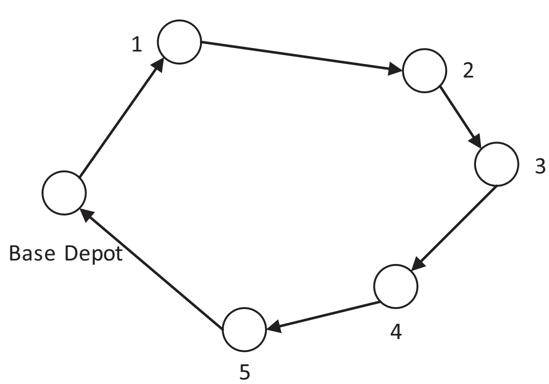


FIGURE 1 Entire closed-loop stop sequence illustrating a trip chain.

Anchor Point Identification

As the starting or anchoring point of a chain, the base depot is important in defining a truck's trip chain. Unfortunately, because of privacy protection, there was no trip purpose information in the raw GPS data to assist in locating the truck depots. However, given the large number of distinct trucks traveling in Washington State and the timestamp information available in the GPS data, it was possible to infer the most frequently visited locations for trucks, which were likely to be their base depots. Those repeatedly visited places are also known as anchor points. In passenger travel pattern studies, anchor points typically refer to home and workplaces and can be found through long-term observation (*31*).

Sharman and Roorda applied the hierarchical clustering (Ward's method) algorithm to find trucks' recurring locations from GPS data (*17*). Their approach is novel and can effectively identify individual trucks' anchor points, but it is computationally intensive and incapable of determining the optimal number of clusters automatically. This study addressed this issue by developing a spatial nonhierarchical clustering algorithm called the density-based spatial clustering of applications with noise (DBSCAN). The algorithm fully incorporates the spatial and temporal features of the GPS data sets and takes advantage of network information. The optimal number of clusters containing anchor points for individual trucks can be determined in an effective and efficient manner.

Geospatial Data Fusion

Trucks' origins and destinations are unlikely to be on freeways; therefore, the GPS records falling on freeways were removed by spatially joining the freeway network data with the GPS latitude and longitude data. By calculating the network distance between each GPS record and freeways, the GPS records in a buffer of less than 10 m from the centerline of each freeway were eliminated. In addition, rest areas and weigh stations are not typically considered as trucks' destinations. A buffer area with a radius of 100 m around these facilities was created to eliminate the relevant GPS records. The operations were implemented in a PostgreSQL database and only required a few Structure Query Language queries to execute in an efficient manner. The size of usable GPS records can be significantly reduced with the use of geospatial data fusion techniques. A total of 67,897 GPS trip ends were eliminated from the total GPS trip O-D data. This led to a 25% processing speed performance gain. This data reduction strategy is particularly useful for the following spatial clustering algorithm.

Identifying Anchor Points with Spatial Clustering Algorithm

The frequency of a truck's arrival at each stop can be explained in two ways. Each truck may visit the same location multiple times within 1 week, or multiple distinct trucks may visit the same location within 1 day. To measure the repetitiveness of the visited stops, spatial and temporal criteria need to be met:

1. The GPS records should be trip ends that have been identified with the previously discussed trip O-D estimation approach.
2. The GPS records should not contain freeways, rest areas, or weigh stations.
3. GPS records that lie within 50 m of each other can be counted as one cluster.
4. The minimum time interval between two consecutive trip ends for each vehicle should be greater than 1 h.

The spatial criteria were based on extensive visual comparisons, and the temporal threshold of 1 h was used to prevent the truck idling issue [i.e., the GPS point may fluctuate (jiggle) when a truck idles] from overestimating the anchor points (*32*). To incorporate these criteria into the clustering procedure, the DBSCAN algorithm (*33*) was applied to group GPS records with similar spatial and temporal characteristics. As opposed to the traditional nonhierarchical clustering algorithms (e.g., *K*-means), the DBSCAN algorithm is able to find arbitrarily shaped clusters and can automatically determine the optimal number of clusters. These features fit especially well with the trucks' anchor point clustering behavior. This is because a base depot where trucks turn off their engines may have a much higher concentration of GPS points than the surrounding area, and the less-dense areas are considered noise, since these places are infrequently visited by trucks.

The concept of the DBSCAN algorithm is straightforward. Two parameters are needed to construct a cluster: Eps and MinPts. Eps regulates the minimum distance necessary for two GPS points to be included in one cluster and was defined as 50 m in this study. MinPts determines the minimum number of GPS points that one cluster should contain. For this study, there should be at least five GPS points in any cluster. The five points were from distinct trucks or the same truck traveling on different days. An assumption was made that one truck may visit its depot for at least 5 weekdays.

The algorithm randomly starts from an unvisited GPS point and searches this point's neighbors that are within the Eps distance to form a cluster. This cluster will expand if nearby points are close enough to the existing points of the cluster and are qualified to be counted (i.e., from the distinct trucks or the same truck on different days). Otherwise, the nearby points will be flagged as noise. This process will continue until no adjacent points can be found. If the number of points within any cluster is smaller than MinPts, this cluster will be discarded.

Average silhouette weights were used to evaluate the effectiveness of the clustering result. By averaging the silhouette weights for all the records within each cluster, the average silhouette weight for the cluster can be calculated between −1 and 1. The clustered results are proven to be more appropriate when the average silhouette weight approaches 1. A sensitivity test was conducted to select the optimal parameter settings for DBSCAN and is shown in Table 1.

The results in Table 1 illustrate that the optimal parameter setting for the DBSCAN was when Eps = 50 and MinPts = 5. The DBSCAN algorithm was then implemented with an indexing structure in Java, and thus the computational complexity was reduced to $O(\log n)$, where n is the total qualified GPS trip ends (182,905 records). As a result, 5,737 clusters containing 104,758 anchor points were generated based on 1 week of data, but the algorithm run time was only 1,292 s. These anchor points were flagged in the databases for further processing.

TABLE 1 DBSCAN Algorithm Parameter Selection

Eps (meters)	MinPts	Number of Clusters	Average Silhouette Weight
25	3	12,153	0.31
50	3	12,141	0.39
100	3	11,077	0.27
25	5	5,819	0.44
50	5	5,737	0.63
100	5	4,805	0.58
25	10	2,117	0.29
50	10	1,852	0.38
100	10	1,332	0.41

Visual Inspection

Satellite imagery provided by Google Earth was used for a visual inspection to confirm the accuracy and reasonableness of the identified anchor points. From the total 5,737 clusters, 100 clusters were randomly selected, as displayed in Figure 2.

Each cluster, which is located around the Puget Sound area in Washington State, contains GPS points identified by the spatial data-mining approach. Individual clusters could be further zoomed in Google Earth and evaluated. Four representative zoom-in clusters with varying geospatial features are demonstrated in Figure 3.

Figure 3*a* shows a depot that is adjacent to the Interstate 5 freeway (the yellow pins are the GPS reads). Any truck traveling on a freeway has been removed with the geospatial data fusion techniques. The GPS points with the status of park are located around the depot building. Figure 3*b* shows a dock where a cargo ship berths. This is considered an important anchor point, since trucks queue up for loading and unloading cargo. The shape of the dock is a narrow strip, and several trucks are stopped next to the cargo ship for discharging. The figure indicates that the DBSCAN algorithm is able to find clusters with arbitrary shapes and can effectively group irregular GPS location data as one cluster.

Figure 3*c* presents a different situation of trucks traveling in a dense urban area. This is challenging to tackle, since different business facilities are located close to each other, and it is difficult to distinguish clusters with traditional clustering algorithms. In Figure 3*c*, the DBSCAN clustering algorithm successfully separated GPS trip ends into three different clusters, although the separate depots are adjacent to each other. The cluster in Figure 3*d* is located at a construction site where several dump trucks queue to load and unload loose material (such as sand, gravel, and dirt) for construction.

To demonstrate further the effectiveness of the proposed clustering algorithm, three graduate students were hired to examine visually the 100 anchor point clusters that were identified in Google Earth. The primary criteria were whether the GPS points within each cluster scattered around a depot-like building where multiple parked trucks can be observed. The validation results reveal that 89 of 100 clusters were reasonably identified by the proposed spatial clustering algorithm. However, 11 clusters were mistakenly detected as actual anchor trip ends.

Two scenarios can be commonly seen from these failed samples, and are demonstrated in Figure 4, *a* and *b*. In Figure 4*a*, the depot covers a large area, with the parking area and docking bays located on both sides of the building. However, because of the small distance threshold (50 m) setting, GPS trip ends on either side of the depot are grouped into two separate clusters. In Figure 4*b*, the GPS

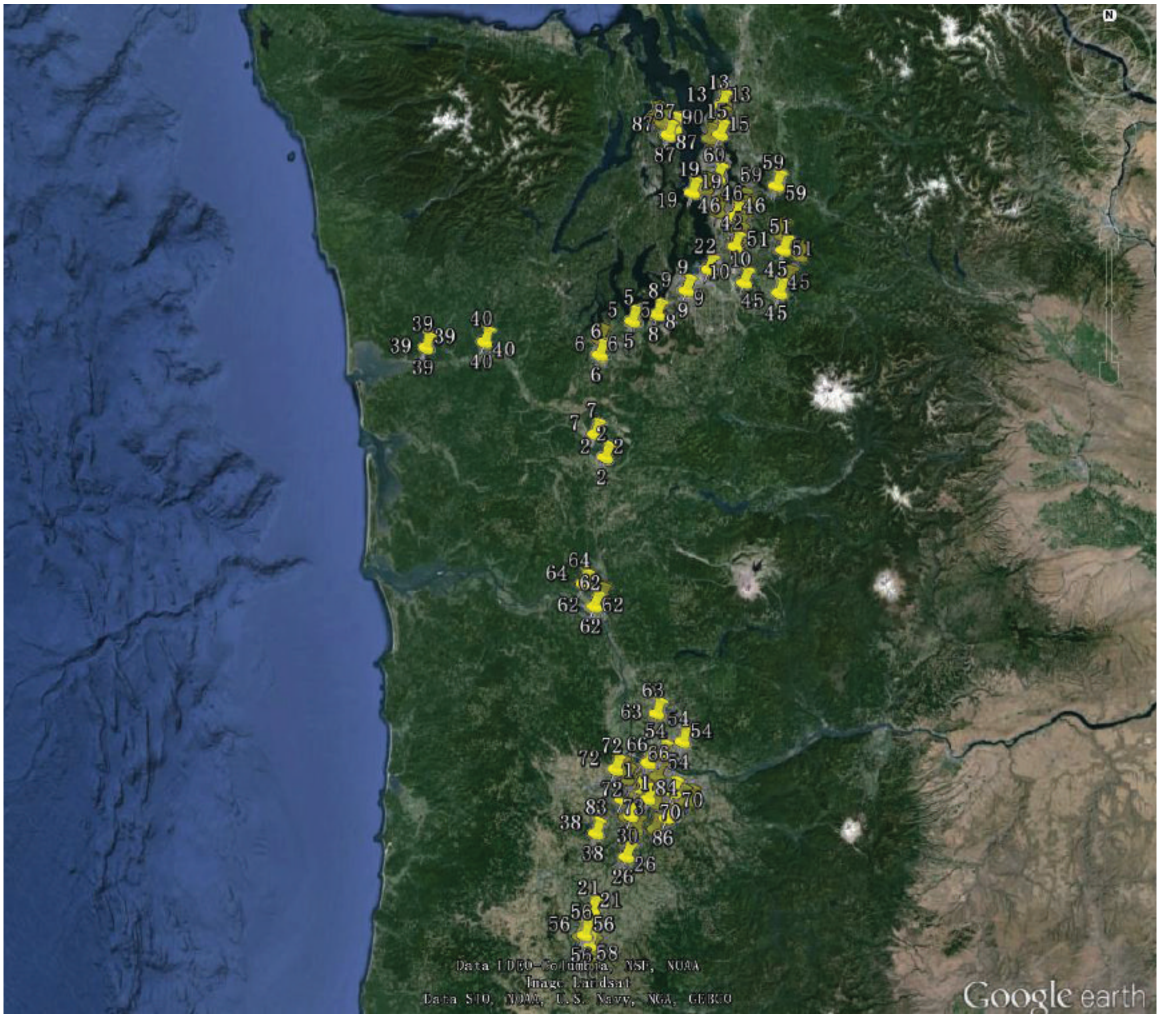


FIGURE 2 Spatial distribution of 100 anchor point clusters.

(a) (b)

FIGURE 3 Sample clustering results with proposed anchor point identification algorithm: (*a*) depot adjacent to Interstate 5 and (*b*) dock where a cargo ship berths.

(continued)

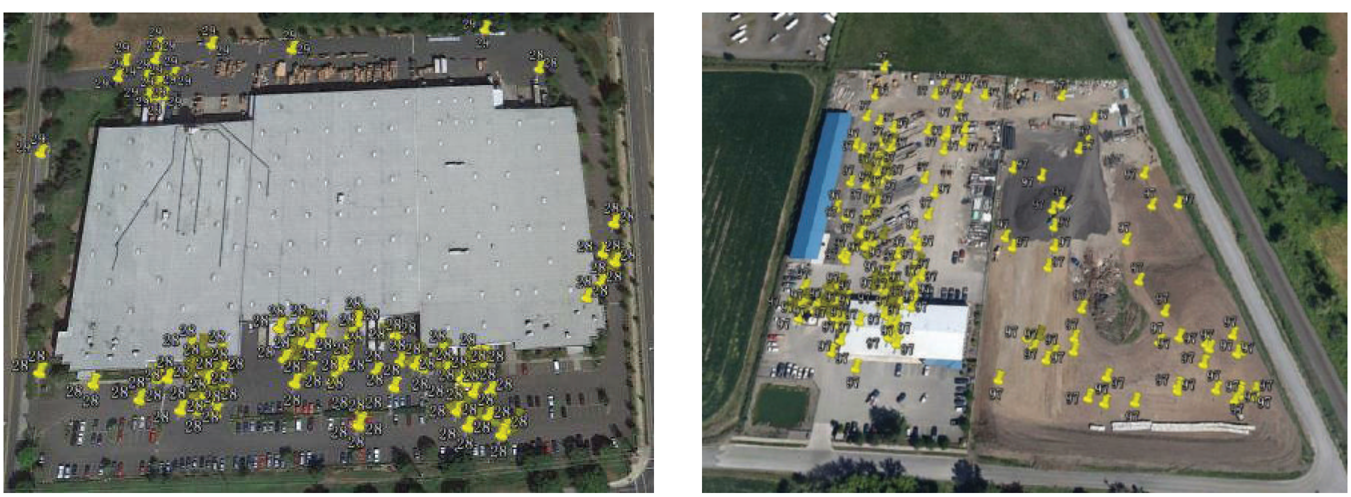
(c) (d)

FIGURE 3 *(continued)* **Sample clustering results with proposed anchor point identification algorithm: (*c*) truck traveling in dense urban area and (*d*) construction site.**

points classified into a single cluster should belong to two clusters, depending on different land use types. However, two different property types coexist within this cluster. The east side of the cluster should be a depot at a large trucking firm, while a construction site is located on the west side of the same cluster. These two facilities are adjoined and are very difficult to separate with a fixed spatial threshold in the DBSCAN algorithm. A possible remedy would be to incorporate detailed land parcel data into the analysis to improve the clustering results (*17*). In this case, the land use attribute of each facility can be taken into account in the spatial clustering procedure.

After the estimated trip ends and anchor points were identified, individual trip chains were constructed. For each truck, the identified trip ends were ordered in a time sequence for each day. The first flagged anchor point of a truck was selected and appointed as the initial point. This anchor point was then sequentially compared with the remaining trip ends until the next anchor point in the same cluster could be found. This process yielded the first trip chain. Meanwhile, travel distance, average dwell time (defined as the time difference between the last trip destination and the current trip origin), as well as number of stops (trip ends) for this trip chain were also calculated. This procedure continued until there were no more anchor points in the same cluster. There could be multiple trip chains that use the same anchor point after this procedure. Similarly, if a second flagged anchor point in another cluster can be found, several additional trip chains can be generated by following these steps. The trip chain construction algorithm will end when there are no more anchor stops for this truck.

Figure 5 shows multiple trip chains for a particular truck with a total of four trip chains for the day. The truck starts from base Depot 1, traverses four intermediate stops, and returns to the base depot as the first trip chain. Then the truck visits other stops from



(a) (b)

FIGURE 4 **Error results for GPS anchor point clustering: (*a*) parking area and docking bays located on both sides of a building and (*b*) two property types within cluster.**

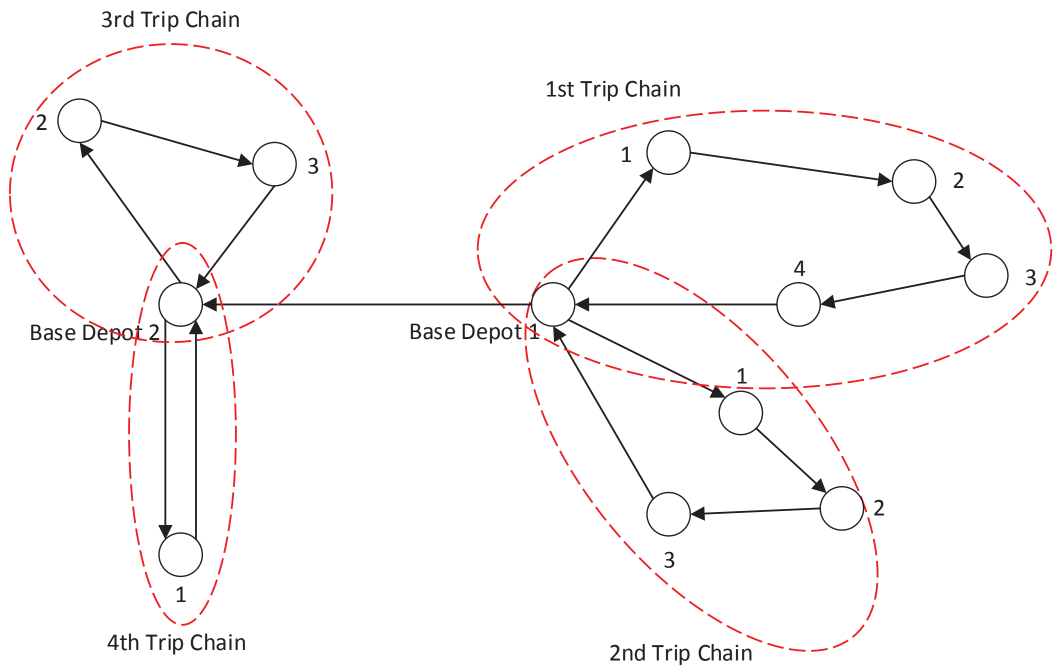


FIGURE 5 Example of multiple trip chains.

base Depot 1, and comes back again. After that, the truck stays at base Depot 1 for a time, and moves to base Depot 2 for multiple deliveries, to form the third trip chain. The truck returns again to base Depot 2 to complete the fourth and final trip chain of the day.

DISCUSSION OF RESULTS AND ANALYSIS

Truck GPS data from May 7 to 9, 2013 (Tuesday to Thursday) were used to generate individual truck daily trip-chaining information with the algorithms. There were 10,425 trucks with 1.3 million GPS records over the 3 days. The fields for each trip-chaining record include truck GPS device ID, number of trip chains, average dwell time, number of stops per trip chain, and average trip length (i.e., average trip distance). The distribution of the trip chains from these days is shown in Table 2.

More than 51% of all the trucks made at least one trip chain during the study week. Among those trucks with at least one trip chain, more than 75% made one or two trip chains per day. This statistic corresponds with the findings of Holguín-Veras and Patil (*12*): as the number of trip chains increases, the average stops per trip chain decreases. A significant number of trucks did not complete a trip chain during a day. In the Washington State study area, there is considerable cross-border traffic with Canada and adjacent states, so it is reasonable to find that some trucks traveled out of state and did not return to their original base depots on the same day. Google Earth was used to follow several trucks without any trip chains, and the results supported this travel pattern.

TABLE 2 Distribution of Trip Chains from May 7 to 9, 2013

Number of Trip Chains	Number of Trucks	Average Stops per Trip Chain	Percentage
0	5,087	na	48.8
1	2,721	7.79	26.1
2	1,303	4.32	12.5
3	584	3.11	5.6
4	251	2.44	2.4
5	156	2.32	1.5
>5	323	1.50	3.1

NOTE: na = not applicable.

Trip-Chaining Behavior Classification

Different categories of trucks have differing chaining behaviors. For instance, long-haul trucks deliver and pick up goods between their home bases, large stores, and warehouses, with few intermediate stops on a daily basis. Small package delivery trucks may complete multiple services per day with short dwell times to drop off packages. These heterogeneous trucking activities are crucial inputs for freight demand modeling, since distinct trip purposes may impact trucks' travel behaviors (*4*). Although several studies have relied on truck survey or diary data to estimate truck trip purposes, little research has been conducted to classify truck categories and travel patterns with GPS data. Because behavioral information does not exist in the raw GPS data, an interesting research question is whether truck category information can be mined with individual trucks' chaining information. This question suggested an approach that uses a nonhierarchical clustering algorithm to categorize individual trucks' overall chaining behavior.

The PAM clustering algorithm was used to group trucks with similar behavioral characteristics. Different from the traditional *K*-means algorithm, the PAM algorithm attempts to minimize the overall dissimilarity between the center of a cluster and its members (*34*). The center of this cluster is named medoids, whose average dissimilarity to all the other points in the same cluster is minimal. This selection of a cluster center ensures that the PAM algorithm has the capability to resist noise and outliers. Given that a total of *n* records should be partitioned into *k* clusters, implementation of the PAM algorithm can be summarized as follows:

Step 1. Randomly choose *k* records as the medoids.
Step 2. Calculate the dissimilarity matrix (Euclidean distance is utilized for simplification) between each record and each medoid, and assign each record to the nearest medoid.
Step 3. For each cluster, search whether there is a record that can lower the average dissimilarity within this cluster. If yes, this record will be the new medoid of this cluster.
Step 4. If the medoids for any clusters are changed in Step 3, repeat Step 2 until there is no change in the medoid.

Four trip-chaining features (average trip chains, average trip stops per trip chain, average dwell time, and average trip distance) were initially selected for clustering. All four features were properly scaled to equalize the contribution of each feature and ensure that the standardized value of each feature fell between 0 and 1. Between 2 and 10 clusters should be selected, and the optimal number of clusters was calculated as four. The trip-chaining statistics for the four clusters are summarized in Table 3.

From Table 3, 53.8% of the total trucks were classified as Cluster 1, and the average number of trip chains for this cluster was the highest, but the average number of stops per trip chain was only 3.28. In comparison with the other clusters, trucks in this cluster only traveled a relatively short distance to complete a trip chain and had a short dwell time at each stop. Therefore, these trucks, in comparison with the total travel distance for all trucks, were more likely to be local delivery trucks transferring goods between their base depots and surrounding retailers or manufacturers.

Cluster 2 included 13.3% of the total trucks, with an average of 1.31 trip chains and about six stops per chain. However, the average dwell time between stops within each trip chain was more than 2 h. This long dwell time further resulted in the shortest average trip chain length (28.66 km). Different possible freight activities at terminals may have contributed to the long delay per stop, such as drayage trucks that had to wait a longer time for cargo to be removed from vessels at a seaport, or delivery trucks that may have been loading or unloading goods at supermarkets for long dwell times.

Cluster 3 was composed of 1,126 trucks with approximately 11 stops per trip chain, and the average duration for each stop was the shortest among all the clusters (12.7 min). On average, each truck completed one trip chain and traveled 101 km per day. The high frequency of stops and short dwell time per stop imply that the trucks in Cluster 3 were likely small package delivery trucks that fulfilled service calls to multiple customers.

The last cluster contained 603 trucks, which was 11.8% of the total trucks. The trucks in Cluster 4 behaved differently from the other trucks. Each truck made about two trip chains per day, and, on average, there were more than four stops for each trip chain. However, the average trip length did not decrease as the number of stops increased. Trucks in Cluster 4 traveled 61.5 km for each trip, which was approximately six times higher than the average trip distance of the trucks in the other clusters. These statistics suggest that the trucks in Cluster 4 were manipulated by local or regional drivers, who work near their homes or only travel within nearby towns for short periods.

Validation

Validating the classified travel patterns was a difficult task because of the lack of detailed travel survey or travel diary data. Fortunately, the GPS traces for a particular truck within each cluster can be visually plotted on Google Earth and used to check the reasonableness of the cluster-based interpretations. Four trucks from the corresponding four clusters were randomly selected, and their waypoints and possible routes were constructed by extracting the 1-day GPS records. A route between two consecutive waypoints was connected by a straight line because of the low level of GPS location update frequency, but each truck's trip-chaining activities can still be recognized easily.

The GPS traces for the randomly selected trucks from Clusters 1 to 4 are respectively demonstrated in Figure 6, *a* to *d*. The truck in Figure 6*a* completed three trip chains for two identified depots. On average, each trip chain was composed of four stops, and associated with 15.5 min of dwell time for each stop. The average travel distance for each trip chain was 52.2 km. The track can be displayed as a span of time on Google Earth. The truck formed the first two trip chains at the beginning of the day. After that, it moved to another depot to serve a sequence of destinations in another area and returned to the depot as the third trip chain at the end of the day.

The truck in Figure 6*b* exhibits unique characteristics, as it left the depot next to a seaport in the morning, visited several destinations, and finally returned to its original location. The evidence supports the assumption that trucks in Cluster 2 experienced long dwell times because they were discharging cargo.

Trucks in Cluster 3 had the most frequent stops within each trip chain compared with the trucks in the other clusters. This is supported by observing the GPS trace of a truck in Figure 6*c*. The truck traveled a long distance (127.1 km) to complete a trip chain of 10 stops, and most of these stops were in residential areas and had a short average dwell time of 6.8 min. These findings imply that it was likely to be a small package delivery truck.

TABLE 3 Trip-Chaining Behavior Classification Results from May 7 to 9, 2013

Cluster Number	Trucks	Average Trip Chains	Average Stops per Trip Chain	Average Dwell Time (min)	Average Trip Chain Length (km)	Average Trip Length (km)
1	2,872	3.28	3.52	16.05	51.45	11.86
2	710	1.31	6.03	138.21	28.66	5.97
3	1,126	1.05	10.98	12.67	101.37	11.44
4	639	1.95	4.41	18.43	244.50	61.53

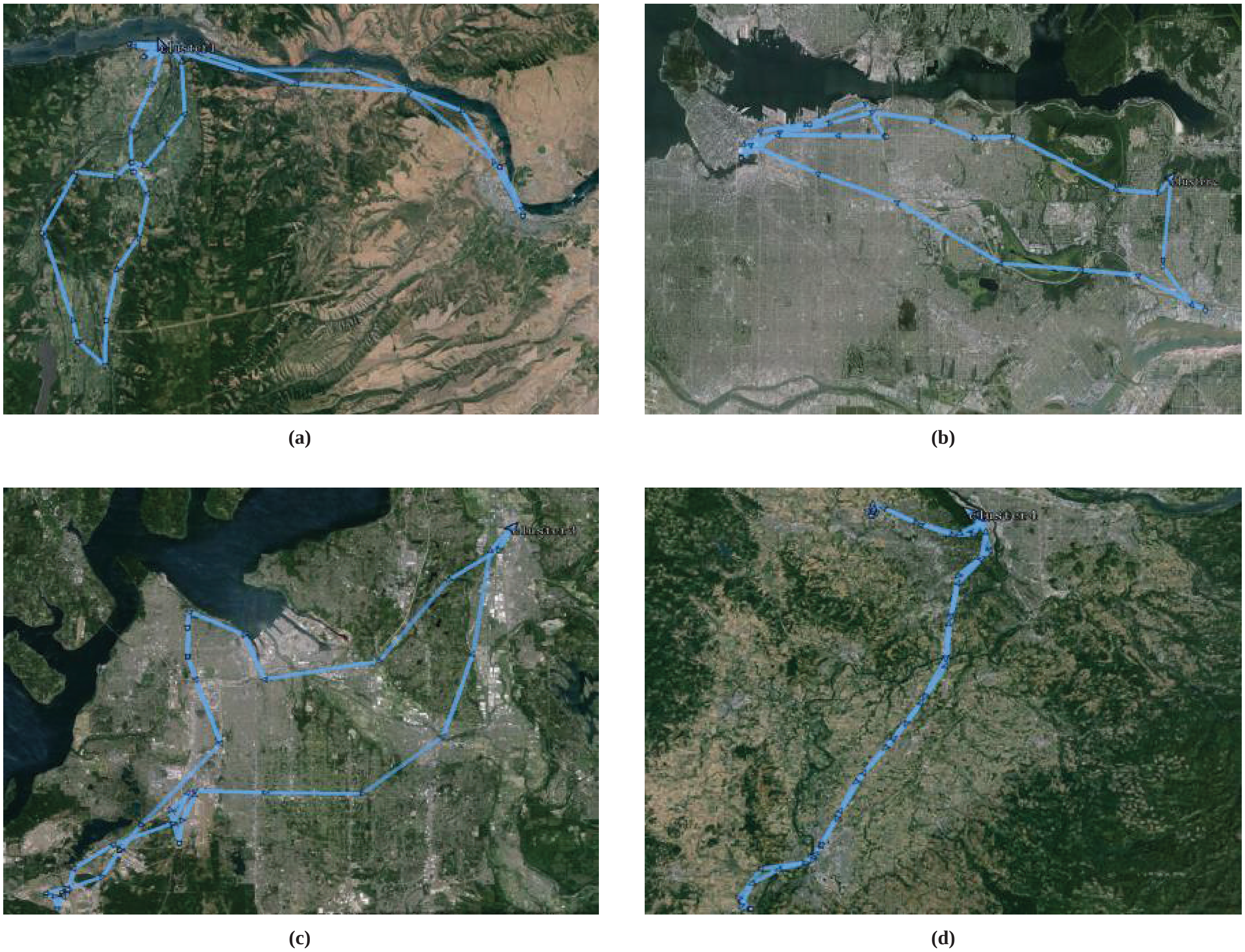


FIGURE 6 GPS traces for a truck in each cluster: (*a*) Cluster 1, (*b*) Cluster 2, (*c*) Cluster 3, and (*d*) Cluster 4.

Figure 6*d* demonstrates a typical truck route for Cluster 4. The truck apparently covered long distances, with a total travel distance of 193.4 km. This long trip chain resulted in only five stops and 18.5 min average dwell time and suggests a regional or local truck driver. Figure 7 further presents the layout of the depot where the truck in Cluster 4 started its trip. An inspection of the truck, trailer size, and terminal layout suggests that long-haul trucks use this location.

Overall, tracking individual trucks' GPS traces within each cluster supports the proposition that the proposed trip chain clustering algorithm is effective.

The freight trip chain extraction and classification methods presented in this paper provide useful information to support freight activity-based model development and freight performance measure monitoring. The identified trip O-D information can be used to track trucks' temporal and spatial movements and can provide travel time, distance, and speed behavior statistics for each O-D pair. Truck trajectories can be visualized by integrating the GPS with a geographic information system (GIS) roadway network. This is particularly useful for understanding truck drivers' route choices to investigate and improve trucking operations. Similarly, the mined freight trip-chaining information can be used to model individual truck drivers' activity and can provide valuable inputs to activity-based models. In addition, the classified truck behavior information is useful for inferring freight trip purpose and further for revealing heterogeneous freight travel patterns to estimate and predict freight demand.

CONCLUSIONS

Freight behavioral research is an important part of freight transportation modeling but is fundamentally different from the common passenger travel behavioral studies. The difference is partially because of the complexity of freight movements, caused by dynamic interactions between various logistics decision makers and the resulting dependencies between freight trips (i.e., trip chaining). Much of the knowledge about freight patterns over the past few decades has been acquired through cross-section data sets, such as travel surveys and travel diaries. As the availability of GPS data increases, there is an opportunity to understand truckers' behavior with the use of data-mining tools and by following digital traces.

Three days of GPS data from 10,425 trucks were processed to generate information for understanding trip-chaining behavior. Approximately 51% of the trucks made at least one trip chain per day, and 49% did not return home within 3 days because they were involved

FIGURE 7 Layout of a truck depot for Cluster 4.

in interstate freight activities (these trucks were likely long-haul trucks). The trucks were classified into categories (clusters) according to their trip-chaining statistics (number of trip chains, number of stops per chain, dwell time, and average trip distance). Four clusters were determined with the PAM nonhierarchical clustering algorithm, and heterogeneous travel patterns were observed within each cluster.

The contributions of this study are twofold. First, the study developed an effective and efficient spatial data-mining approach to extract trip-chaining information from large GPS data sets. Second, as opposed to traditional survey methods, the generated trip-chaining data can be expanded into multiple days to understand freight travel behavior dynamics and eventually facilitate the development of freight demand models. Knowing the heterogeneity of freight trip-chaining behaviors allows for fine-tuning freight activity forecasting models. This can be done through agent-based simulation platforms to mimic freight movements by incorporating the trip-chaining features (number of trips per chain, dwell time, average trip length, and so forth).

Although the proposed spatial data-mining and clustering algorithms are promising as a potential approach to understand freight behavior, GPS-based data collection has inherent disadvantages in the sample size issue. Although the number of GPS equipped trucks is growing, it still only represents a nonrandomly selected subset of total truck activities within a spatial area. Therefore, it may be necessary to incorporate traditional freight data collection methods, such as travel surveys and diaries, to lower the sampling errors, or verify the uniform market penetration rate assumption to estimate the unbiased freight O-D demands (*34*).

Nevertheless, as MAP-21 requires that state DOTs develop or improve existing freight data collection methods for performance-based transportation planning and programming, it can be seen that automatic data collection methods will be a trend to support the development of freight performance measures. The market penetration rate of GPS devices will increase, and this will reduce sampling problems. In addition, survey data should never be neglected for extracting freight behavioral information. A driver's perception and route preference information cannot be captured with GPS data. By combining survey data and GPS data, model parameters can be better tuned up and the validation results can be more persuasive and trustable.

Further research could enhance the ideas developed in this study. The trip-chaining behavior classification results could be compared with the ground-truth trip purposes or truck category data from freight surveys or travel logs. In addition, the accuracy of the depot anchor point identification algorithm could be further improved by integrating the depot location with land use data in a GIS. This study also opens new possibilities for investigating long-term freight trip chain regularity and variability with the use of multiday GPS data.

ACKNOWLEDGMENTS

This research was supported by the National Natural Science Foundation of China; National Key Technologies R&D Program of China; Beijing Nova Programs; Natural Science Foundation of Chongqing, China; and Fundamental Research Funds for the Central Universities.

REFERENCES

1. *Pocket Guide to Transportation in America: A New National Picture.* Research and Innovative Technology Administration, Bureau of Transportation Statistics, U.S. Department of Transportation, 2013. http://www.rita.dot.gov/bts/publications/pocket_guide_to_transportation/2013. Accessed Oct. 21, 2013.
2. Wang, Y., X. Ma, Y. Lao, and Y. Wang. A Fuzzy-Based Customer Clustering Approach with Hierarchical Structure for Logistics Network Optimization. *Expert Systems with Applications,* Vol. 42, No. 2, 2014, pp. 521–534.
3. *Moving Ahead for Progress in the 21st Century Act (MAP-21).* FHWA, U.S. Department of Transportation, 2012. https://www.fhwa.dot.gov/map21/docs/map21_summary_hgwy_provisions.pdf. Accessed July 28, 2013.
4. Ruan, M., J. Lin, and K. Kawamura. Modeling Urban Commercial Vehicle Daily Tour Chaining. *Transportation Research Part E,* Vol. 48, No. 6, 2012, pp. 1169–1184.

5. Holguín-Veras, J., and E. Thorson. Trip Length Distributions in Commodity-Based and Trip-Based Freight Demand Modeling: Investigation of Relationships. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1707,* TRB, National Research Council, Washington, D.C., 2005, pp. 37–48.
6. Adler, T., and M. Ben-Akiva. A Theoretical and Empirical Model of Trip-Chaining Behavior. *Transportation Research Part B,* Vol. 13, No. 3, 1979, pp. 243–257.
7. Axhausen, K., and T. Gärling. Activity-Based Approaches to Travel Analysis: Conceptual Frameworks, Models and Research Problems. *Transport Reviews,* Vol. 12, No. 4, 1992, pp. 324–341.
8. Misra, R., C.R. Bhat, and S. Srinivasan. Continuous Time Representation and Modeling Framework for Analysis of Nonworker Activity-Travel Patterns: Tour and Episode Attributes. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1813,* Transportation Research Board of the National Academies, Washington, D.C., 2003, pp. 11–20.
9. Kitamura, P. Trip-Chaining in a Linear City. *Transportation Research Part A,* Vol. 19, No. 2, 1985, pp. 155–167.
10. Golob, T. F. A Nonlinear Canonical Correlation Analysis of Weekly Trip Chaining Behavior. *Transportation Research Part A,* Vol. 20, No. 5, 1986, pp. 385–399.
11. Bhat, C.R., and F.S. Koppelman. Activity-Based Modeling of Travel Demand. In *The Handbook of Transportation Science* (R.W. Hall, ed.), Kluwer Academic Publishers, Norwell, Mass., 1999, pp. 35–61.
12. Holguín-Veras, J., and G.R. Patil. Observed Trip Chain Behavior of Commercial Vehicles. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1906,* Transportation Research Board of the National Academies, Washington, D.C., 2005, pp. 74–80.
13. McGuckin, N., and Y. Nakamoto. Trips, Chains, and Tours: Using an Operational Definition. Presented at Understanding Our Nation's Travel: National Household Travel Survey Conference, Washington, D.C., 2004.
14. Wang, Q., and J. Holguín-Veras. Investigation of Attributes Determining Trip Chaining Behavior in Hybrid Microsimulation Urban Freight Models. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2066,* Transportation Research Board of the National Academies, Washington, D.C., 2008, pp. 1–8.
15. Samimi, A., A. Mohammadian, and K. Kawamura. A Behavioral Freight Movement Microsimulation Model: Method and Data. *Transportation Letters,* Vol. 2, No. 1, 2010, pp. 53–62.
16. Figliozzi, M.A. Analysis of the Efficiency of Urban Commercial Vehicle Tours: Data Collection, Methodology, and Policy Implications. *Transportation Research Part B,* Vol. 41, No. 9, 2007, pp. 1014–1032.
17. Sharman, B.W., and M.J. Roorda. Analysis of Freight Global Positioning System Data: Clustering Approach for Identifying Trip Destinations. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2246,* Transportation Research Board of the National Academies, Washington, D.C., 2011, pp. 83–91.
18. Ma, X., E.D. McCormack, and Y. Wang. Processing Commercial Global Positioning System Data to Develop a Web-Based Truck Performance Measures Program. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2246,* Transportation Research Board of the National Academies, Washington, D.C., 2011, pp. 92–100.
19. Fagerberg, J. Fleet Management-Word. 2012. http://www.marketresearch.com/Berg-Insight-v2702/Fleet-Management-7095368/. Accessed Oct. 21, 2013.
20. Greaves, S.P., and M.A. Figliozzi. Collecting Commercial Vehicle Tour Data with Passive Global Positioning System Technology: Issues and Potential Applications. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2049,* Transportation Research Board of the National Academies, Washington, D.C., 2008, pp. 158–166.
21. Hensher, D., and M.A. Figliozzi. Behavioral Insights into the Modeling of Freight Transportation and Distribution Systems. *Transportation Research Part B,* Vol. 41, No. 9, 2007, pp. 921–923.
22. Wang, Q. *Tour-Based Urban Freight Travel Demand Models.* PhD dissertation. Rensselaer Polytechnic Institute, Troy, N.Y., 2008.
23. Hunt, J., and K. Stefan. Tour-Based Microsimulation of Urban Commercial Movements. *Transportation Research Part B: Methodological,* Vol. 41, No. 9, 2007, pp. 981–1013.
24. Figliozzi, M.A. The Impacts of Congestion on Commercial Vehicle Tour Characteristics and Costs. *Transportation Research Part E,* Vol. 46, No. 4, 2010, pp. 496–506.
25. Li, H., R. Guensler, J. Ogle, and J. Wang. Using Global Positioning System Data to Understand Day-to-Day Dynamic of Morning Commute Behavior. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1895,* Transportation Research Board of the National Academies, Washington, D.C., 2004, pp. 78–84.
26. Du, J., and L. Aultman-Hall. Increasing the Accuracy of Trip Rate Information from Passive Multi-Day GPS Travel Datasets: Automatic Trip End Identification Issues. *Transportation Research Part A,* Vol. 41, No. 3, 2007, pp. 220–232.
27. Alvarez-Garcia, J.A., J.A. Ortega, L. Gonzalez-Abril, and F. Velasco. Trip Destination Prediction Based on Past GPS Log Using a Hidden Markov Model. *Expert Systems with Applications,* Vol. 37, No. 12, 2010, pp. 8166–8171.
28. Spissu, E., I. Meloni, and B. Sanjust. Behavioral Analysis of Choice of Daily Route with Data from Global Positioning System. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2230,* Transportation Research Board of the National Academies, Washington, D.C., 2006, pp. 96–103.
29. McCormack, E.D., and W. Zhao. GPS Truck Data Performance Measure Program in Washington State. Research Report Prepared for the Washington State Department of Transportation, Olympia, 2011. http://www.wsdot.wa.gov/NR/rdonlyres/B7A5D60C-BF99-412F-9444-BA513768DC93/0/TPMFinalreportver2_17June2011.pdf. Accessed Oct. 21, 2013.
30. McCormack, E.D., and M.E. Hallenbeck. ITS Devices Used to Collect Truck Data for Performance Benchmarks. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1957,* Transportation Research Board of the National Academies, Washington, D.C., 2006, pp. 43–50.
31. Ma, X., Y. Wu, Y. Wang, F. Chen, and J. Liu. Mining Smart Card Data for Transit Riders' Travel Patterns. *Transportation Research Part C,* Vol. 36, 2013, pp. 1–12.
32. Tang, J., F. Liu, Y. Wang, and H. Wang. Uncovering Urban Human Mobility from Large Scale Taxi GPS Data. *Physica A,* Vol. 438, No. 15, 2015, pp. 140–153.
33. Ester, M., H.P. Kriegel, J. Sander, and X. Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining.* AAAI Press, 1996, pp. 226–231.
34. Wang, Y., X. Ma, M. Xu, Y. Liu, and Y. Wang. Two-Echelon Logistics Distribution Region Partitioning Problem Based on a Hybrid Particle Swarm Optimization-Genetic Algorithm. *Expert Systems with Applications,* Vol. 42, No. 12, 2015, pp. 5019–5031.

The Standing Committee on Freight Transportation Data peer-reviewed this paper.