Deriving Truck Route Choice from Large GPS Data Streams

Mohammadreza Kamali
Department of Civil and Environmental Engineering
University of South Florida
4202 E Fowler Avenue, ENC 3300, Tampa, FL 33620
Tel: +1 (813) 713-1327; E-mail: mkamali@mail.usf.edu

Alireza Ermagun
Department of Civil, Environmental, and Geo- Engineering
University of Minnesota, Twin Cities
500 Pillsbury Drive SE, Minneapolis, MN 55455
Tel: +1 (612) 701-0440; Email: ermag001@umn.edu

Krishnan Viswanathan
Cambridge Systematics, Inc.
1566 Village Square Blvd, Suite 2, Tallahassee FL 32309
Tel: +1 (850) 671-0203; E-mailkviswanathan@camsys.com

Abdul R. Pinjari, Ph.D. (*Corresponding Author*)
Department of Civil and Environmental Engineering
University of South Florida
4202 E Fowler Avenue, ENC 3300, Tampa, FL 33620
Tel: +1 (813) 974-9671; E-mail: apinjari@usf.edu

Word Count (excluded references): 5,787 (Text) + 1,500 (2 Tables + 4 Figures) = 7,287 words

TRR Paper number: 16-4095

Submission Date: 3/15/2016

ABSTRACT

This study develops a simple and effective procedure to derive truck route choice data using large streams of truck GPS data. At the heart of this procedure is a map-matching algorithm that is easy to implement. The proposed procedure is applicable to large datasets with varying frequencies between consecutive GPS records, ranging anywhere from as high frequency as one time per second to as low frequency as one time in 20 minutes. As a demonstration, the procedure is employed to derive routes for over 78,000 truck trips resulting from a large GPS dataset. The derived routes are validated to demonstrate the accuracy of the results. We also introduce a metric useful for measuring the amount of overlap between the travel path of one trip to that of another trip between a given origin-destination (OD) pair. In addition, we develop a procedure to employ the metric for measuring the diversity (or variability) of routes across several trips between a given OD pair and to identify the set of unique routes used between the OD pair.

INTRODUCTION

Freight mobility is critical to sustain the anticipated growth in trade and freight movement in the United States. Currently, around 70 percent of all commodities shipped (by weight) is via trucks *(1)*. Dominance of the truck mode is expected to continue and contribute to congestion and wear-and-tear of the infrastructure. Traffic congestion hinders freight mobility on highways. An essential step toward enhancing highway freight mobility is to gain a thorough understanding of freight truck travel behavior, including the demand for travel between origins and destinations, interaction with other modes of travel, and the routes of travel. An undeveloped dimension of these aspects is truck travel route choice. It is essential to understand and forecast truck travel route choice, and the aggregate level network performance for medium- to long-term decisions such as the designation of truck routes, addition of new truck corridors and by-pass routes, and resource allocation for infrastructure maintenance.

A primary challenge for all such investigations is the lack of observed data on truck routes. Traditional travel surveys do not allow for the observation and measurement of truck travel routes. One way to overcome the data related challenges in observing and measuring truck travel routes is to utilize advanced vehicle monitoring (AVM) systems that allow remote monitoring of truck fleets using Geographical Positioning Systems (GPS) technology. The increasing availability of GPS data opens significant opportunities for visualization as well as accurate measurements of truck travel routes. Further, access to large amounts of truck GPS data offers unprecedented opportunities to examine the diversity of truck travel routes for any given origin-destination (OD) pair of interest. While typical highway networks present a large number of route choice alternatives between any given OD pair, trucking companies and/or truck drivers may not consider all available options when making the route choice. In this context, using GPS data to analyze the diversity of observed travel routes might help in devising choice set generation algorithms for route choice modeling.

An important precursor step for most analyses and modeling of truck routes using GPS data is to derive the exact travel routes traversed by trucks on the network. Any quantitative analysis and modeling of routes requires the measurement of routes in the form of all network links and nodes traversed by the vehicles between the trip origin and destination. While mapping the GPS data points of travel between the origin and destination of a trip in a GIS environment helps in visualizing the travel route, such data alone may not suffice for measurement of the entire route (i.e., all highway links and nodes) traversed between the origin and destination. It is important to first map the GPS data to the network and then derive the full travel path between the origin and destination. This motivates the need for simple yet effective map matching procedures to derive travel routes using truck GPS data.

The first objective of this study therefore is to develop a simple yet effective map-matching approach for generating truck routes from large streams of truck GPS data. The framework is designed to be applicable to a large amount of data with varying frequencies (i.e., varying time intervals between consecutive GPS records) ranging anywhere from as high frequency as a second to as low frequency as 20 minutes. As a demonstration, the framework developed in this research is used to derive accurate routes for over 78,000 truck trips developed from a large GPS dataset (over 100 Million GPS data points) obtained from the American Transportation Research Institute (ATRI). The derived routes are also validated to demonstrate the accuracy of the results. The second objective is to develop metrics for measuring the diversity (or dissimilarity or variability) of routes across several trips between a given OD pair and to identify the set of unique routes used between an OD pair. The usefulness of the proposed metric is demonstrated by identifying the number (and diversity) of unique truck routes between several OD pairs.

The remainder of the study is organized as follows: We present a brief review of the literature relevant to map matching of GPS data for deriving travel routes. We then describe the GPS data used for the current research, followed by presenting the procedure developed to derive the truck travel routes along with the results of applying the procedure to the GPS data. Finally, the study concludes with a discussion of the methods and results, its applicability for model development, and recommendations for further research.

LITERATURE REVIEW

Map matching is a technique that uses a combination of GPS location data and roadway network data to identify the correct link that has been traversed by the vehicle on the network. The first documented technique of map matching dates to 1996 when Kim et al. *(2)* introduced a simple algorithm that mapped the GPS points to the closest node or shape point in the network. Since then, a variety of map matching algorithms have been developed, which can be grouped into the following categories: geometric based, geometric and topological based, and probabilistic based algorithms. The geometric based algorithm uses the distance of either point-to-curve, curve-to-curve, or the angle of curve-to-curve for map matching. The geometric and topological based approach attempts to reduce incorrect map-matching from a purely geometric approach by considering the connectivity of the network elements. The probabilistic based algorithm defines a confidence region that is defined around each GPS point which is then imposed on the road network to understand the road segments that are part of the route based on closeness, connectivity, and heading *(3)*. The current map matching algorithms in literature have several limitations that inhibit their applicability to large streams of less frequent GPS data. First, several methods that are useful only for high frequency GPS data may not provide accurate routes when used for sparsely spaced data. Second, most applications in the literature (particularly those that used advanced methods) are for small to moderate sized datasets (i.e., for deriving routes of at most a few hundred trips) and are difficult to apply for large streams of GPS data. Third, most previous map matching applications are in the context of dense urban networks, as opposed to data from long-haul trucks that usually traverse major highways and arterials.

Methods to validate these map matching algorithms may be grouped under the following categories: site based methods, comparison methods, and analytical methods. Site based methods rely on a vehicle traversing a route and comparing it to a route chosen via map matching. However, this method is not practical when considering large study areas such as an entire state. Comparison methods, as the name implies, compares the map matching results with an already validated map-matched data. While comparison methods overcome some of the limitations of site based methods, they demand either larger datasets or already validated data. Analytical methods such as feasibility and continuity analysis done by Hess et al. *(4)*, and correct road matching ratio implemented by Jagadeesh et al. *(5)* are more cost effective, but they still need an already validated dataset to serve as the basis of the comparison.

DESCRIPTION OF THE GPS DATA

The GPS data used for this research was obtained from ATRI for the purpose of deriving statewide OD flows of truck movements in the state of Florida for four months – March, April, May, and June – in 2010 [see Pinjari et al. *(6)* and Zanjani et al. *(7)*]. Specifically, for each of these four months, all trucks from ATRI’s database that were in Florida at any time during the month were extracted. Subsequently, all GPS records of those trucks were extracted for the entire month, as they traveled within Florida as well as in other parts of North America. The number of GPS records for each month was over 35 million, summing up to over 145 million records for the four months.

As described in Pinjari et al. *(6)*, each GPS record in ATRI's database contained information on its spatial (latitude/longitude) and temporal (date/time) location along with a unique truck ID that did not change across all the GPS records of the truck for a certain time period (at least two weeks for most trucks in the data). A portion of the GPS data contained spot speeds (i.e., the instantaneous speeds) of the truck and the remaining portion did not contain spot speeds. These two types of data were separately delivered, presumably because they come from different truck fleets with different GPS technologies. The frequency of the GPS data streams varied considerably, ranging from a few seconds to over an hour of intervals between consecutive records. Similarly, the spatial gap between consecutive coordinates in the data varied anywhere from zero to a few feet to several miles. Furthermore, for the year 2010, ATRI's truck GPS data with spot speeds contained higher frequency GPS data streams than ATRI's GPS data without spot speeds.

PROCEDURE TO GENERATE ROUTES FROM RAW GPS DATA

The focus of this section is on the process of generating routes from raw GPS data streams and validating the derived routes.

Conversion of Raw GPS Data into Trips

As a first step, the raw GPS data needed to be converted to truck trips in order to be ready for route generation. The process is briefly summarized below, but explained in detail by Thakur et al. *(8)*:

1. Sort GPS data for each truck ID in the order of date & time of GPS records.
2. Identify an initial set of truck trip stops (i.e., trip ends) based on spatial movement, time gap, and speed between consecutive GPS points. A truck is considered to have stopped at a destination if it stops (i.e., if the average travel speed between two consecutive GPS points is less than 5 mph) for at least 5 minutes. A truck stop of less than 5-minute duration is considered to be a traffic stop (i.e., not a valid destination) and, therefore, considered a part of the travel between origin and destination.
3. Eliminate stops in rest areas as these are unlikely to be pickup/delivery stops:
    - By overlaying trip ends on a GIS file of rest areas and wayside parking stops.
    - By eliminating stops within close proximity of interstate highways, most of which are most likely to be rest areas or wayside parking stops.
    - By joining consecutive trips ending and beginning at such stops.
4. Eliminate circuitous trips (with ratio of airline distance to geodetic distance < 0.7)

Map-Matching Dataset Preparation

Following the conversion of raw GPS data into truck trips, the data was further refined to obtain a dataset ready for map-matching of the GPS points to an underlying highway network. Two broad stages in this process are discussed below.

Selection of trips suitable for accurately deriving routes

The spatial proximity between consecutive coordinates in the GPS data plays an important role in the accuracy of the routes generated using the data. The larger the spatial gaps between consecutive coordinates, derived route is likely to be less accurate. Since spatial separation in the data is dependent on the temporal frequency of the data, it is likely that higher frequency datasets provide more accurate measurement of routes. And recall from earlier discussion that ATRI’s truck GPS data with spot speeds has higher frequency data streams than their data without spot speeds. Therefore, the trips derived from ATRI’s truck GPS data are separated into two categories: (a) trips derived from the data with spot speeds, and (b) trips derived from the data without spot speeds. For both these types of trips, the raw GPS data was used to analyze the temporal and spatial gaps (between consecutive coordinates) in the data. Tables 1 and 2 show the cross-tabulation between the largest time gap (between any two consecutive coordinates) for a trip and the corresponding spatial gap (i.e., for the same pair of consecutive coordinates for which largest time gap was observed in a trip). The reason we analyze the largest time gap and corresponding spatial gaps for each trip is to examine the worst case scenario for each trip (the larger the gap between consecutive coordinates the more difficult it is to map-match to a network).

Table 1 shows that 70 percent of all the trips derived from the data with spot speeds have the largest time gap of at most 20 minutes and the corresponding spatial gap of at most 20 miles. Table 2 shows that 44 percent of all the trips derived from the data without spot speeds have the largest time gap of at least 45 minutes and a corresponding spatial gap of at least 30 miles. These results suggest the trips derived from the data without spot speeds have very sparse (in time and space) GPS data. It is difficult to derive accurate travel route information from such sparse GPS data. Therefore, further analysis focused on only those trips derived from the data with spot speeds. Among the trips derived from the data with spot speeds, the following two categories of trips are kept: (a) trips that have the largest time gap of at most 20 minutes and the corresponding spatial gaps of at most 20 miles, and (b) trips with largest time gap beyond 20 minutes but with the corresponding spatial gaps less than 5 miles. The second category of trips was retained despite high time gaps because the corresponding spatial gaps were relatively small suggesting very slow movements in a large time frame. In summary, over 97 percent of the trips derived from the data with spot speeds were retained for further analysis. The corresponding cells in Table 1 are highlighted in grey color. Furthermore, any remaining trips with the largest spatial gap (between any consecutive coordinates) of more than 20 miles and travel speed of over 100 mph between any consecutive GPS coordinates are eliminated from the dataset. Such data with unrealistically large travel speeds between consecutive GPS data points is most likely due to GPS errors.

Insert TABLE 1 here

Insert TABLE 2 here

The above discussed criteria were devised to select trips suitable for accurately deriving routes based on the temporal and spatial proximity and accuracy of the GPS data. The following additional criteria were then employed to focus the analysis on specific types of trips or OD pairs.

1. Selected only those trips that started and ended in Florida since the research team has access to a detailed highway network (from Navteq) only for the state of Florida;
2. Selected only those trips that did not start and end in the same Traffic Analysis Zone (TAZ);
3. Selected only those trips belonging to OD pairs that had at least 20 trips derived from the data. For analyzing variability of route choices between any given OD pair, observed route data of at least 20 trips between that OD pair were deemed to be sufficient; and
4. Selected only those trips of length greater than 5 miles.

Preparation of the GPS data for map-matching

For all the remaining trips after the above discussed refinements, the corresponding raw GPS data has to be prepared to get it ready for map-matching. To do so, the following four steps were employed on the raw GPS data of each trip.

1. Step 1: Sample GPS coordinates at every 5 minutes (or more, based on the time gaps in the raw data). This was done to reduce the computation costs of map-matching all raw GPS data points. As will be shown later, such time-sampling did not reduce the accuracy of the routes derived.
2. Step 2: Remove GPS points within one-mile radius of origin/destination for each trip. For trips that started and/or ended in urban areas with high-density of highway network links, it is not easy to accurately map-match the raw GPS data. This is because the study highway network data, as is typical with such network datasets, are not detailed enough to represent all minor roadway links at the trip ends. Unlike previous studies that used very high frequency GPS data and without any variations in the time gaps and spatial gaps, map-matching with the current data was more prone to inaccuracies (and resulted in loops or irrational circular moves in the derived routes) near the trip ends. Removing GPS points within one-mile buffer around origin/destination helps avoid such loops. Besides, accurately deriving the routes of short, back-and-forth movements made by trucks near the trip ends (for example, for finding a parking space) are not necessarily of interest in the context of long-haul trips made by trucks. This step also removes origin and destination GPS points. These O and D points are later added at the last step of the map-matching algorithm.
3. Step 3: Remove GPS points with spot speeds less than 20 mph. This helps in eliminating situations when the truck reduced its speed to stop during the trip. Including such stops in the data would cause detours in the generated routes. The idea is to retain only the GPS data points where the truck was moving fast enough to avoid detours or loops in the derived routes.
4. Step 4: Remove trips that had less than 3 GPS points remaining after all the above steps. Such trips were removed because all the previous steps reduced the GPS data below the amount needed to accurately derive the travel route.

After implementing the above steps, a database of 1.53 Million truck trips derived from over 53 Million raw GPS points with spot speeds was reduced to a total of 84,236 trips (corresponding to 725,483 GPS points) for subsequent map-matching and route determination.

The Proposed Map-Matching Procedure

The proposed map-matching procedure is a modified version of an algorithm introduced by Yang et al *(9)*, as described below:

- Step 1: Overlay the GPS data points on the highway network. Use ArcGIS Network Analysist tool to find the closest and second closest highway network links to each GPS point. To map-match a given GPS point, let D1 and D2 denote the distances from its spatial location to the closest network link and the second closest network link, respectively.
- Step 2: If D1 > 1,000 ft. then remove the GPS point. GPS points that have no network links within their 1000 ft. buffer are difficult to map-match. This step eliminates such GPS points to avoid matching them to the wrong link.
- Step 3: If $\frac{D2}{D1} > 2$ then go to *"step 4"*; or else, go to *"step 5"*.
- Step 4: If D1 + D2 > 35 ft. then match the GPS point to the closest network link. Otherwise, remove the GPS point. This step has been implemented to avoid matching GPS points to the wrong link at interchanges or near ramps. Since network links are very close to each other at such places, D1 and D2 might be smaller than maximum spatial accuracy of GPS locations. This might lead to matching the GPS point to a wrong link. Therefore, there should be a lower bound on D1+D2 to make sure D1+D2 is greater than twice of GPS maximum accuracy. This lower bound was set to 35 ft. because maximum spatial accuracy of typical GPS datasets is 5 meters (16.4 ft.) *(10)*. Yang et al. *(9)* do not consider this step, but we implement it to consider (in)accuracies in typical GPS datasets.
- Step 5: Make a 65 ft. buffer around each GPS point that did not satisfy the ratio in *"step 3"*. If there is only one intersection node falling within this buffer, then match the point to the intersection node. This indicates situations where the GPS point is near a single intersection node; and matching it to the node (were both links meet) resolves the issue of which link to match it to. On the other hand, if the buffer does not contain any intersections or contains two or more intersection nodes, then remove the GPS point from the dataset. In both these situations, we considered it prudent to remove the GPS data point than to utilize complex procedures to match the GPS point to the network. Situations with no intersections within a 65ft. buffer indicate that the GPS point is close to two non-intersecting links. We removed such GPS data points, unlike Yang et al. *(9)*, because matching such GPS points to an incorrect link can potentially lead to significant inaccuracies in the derived routes. And situations with multiple intersections inside the buffer indicate intersections that are coded as multiple nodes in the network data. Since it is difficult to decide the node to which the GPS point should be matched to, it was decided to remove GPS points that had two or more intersection nodes falling inside their buffers.
- Step 6: Add the origin and destination GPS coordinates to the set of GPS records for each trip.
- Step 7: Remove any trip that has less than 5 GPS points (including origin and destination coordinates) after all the above described steps. This is because trips with fewer GPS data points may not provide sufficient information to derive the travel route.

After the map-matching process was implemented the final dataset had a total of 78,381 trips, for which travel routes are generated in the next step. This map-matching algorithm provides a reasonably high level of accuracy (as will be demonstrated later) while also being easy to implement for large streams of GPS data. Most map-matching methods that might result in high accuracy outputs utilize complicated algorithms that may not be easy to implement in addition to being computationally intensive. Further, most applications of such algorithms have been on small-sized datasets. In addition, the implementation software of such advanced algorithms are not available in the public domain. The proposed method in this study benefits from a very simple procedure that can easily handle large GPS datasets while maintaining a satisfactory level of accuracy. Equally important, it can be implemented using widely used software packages such as ArcMap thereby helping reach a wider audience.

Route Generation

Due to the infrequent nature of the GPS data in this study, the map-matching algorithm described above does not reveal all roadway links traversed by a truck on its trip. Only those links that corresponded to the GPS data are revealed. Therefore, all missing links need to be assigned so that a complete route may be generated for each trip. To this end, ArcMap 10.3 Network Analyst extension was employed. Specifically, for each trip, the entire travel route was generated by deriving the shortest time path between the origin and destination and through the intermediate, map-matched roadway links between the origin and destination. Network Analyst utilizes a modified version of Dijkstra's algorithm to find shortest paths between two consecutive GPS points map-matched to the network.

The final output of route generation is a GIS shapefile in which each feature is a network link that contains network information as well as trip information. Figure 1, in its left panel, shows an overall view of the routes generated for 78,381 trips between 2,237 OD pairs in Florida. In the same figure, the right panel shows all the generated routes for 218 trips between one specific OD pair. In this example the origin TAZ is in Polk County (in central Florida) and the destination TAZ is in Miami-Dade County (in southeast Florida).

Insert FIGURE 1 here

Validation of the Generated Routes

The generated routes were validated in terms of consistency and feasibility. Specifically, a route is considered to be consistent if the direction of its travel is consistent throughout the entire route and there are no loops in the route, and if there are no infeasible maneuvers (such as jumping of a bridge) throughout the entire route. To check consistency, each generated route was compared to the route from Google Earth to determine if the generated route shows the same direction for most part of the trip. For example if the truck has to take the north bound direction on the highway to get to the destination, the generated route should show that the truck has maintained that direction for a large part of the trip. For feasibility check, each trip is observed at interchanges/overpasses/ramp junctions to see if the generated route shows any impossible maneuvers at such locations.

A random sample of 80 trips was selected for the validation exercise. The generated route for each of these trips was followed on Google Earth to verify its consistency and feasibility simultaneously. Figure 2 shows an example of consistency and feasibility check for one OD pair. The arrows show the direction of travel depicted by the generated route. It is consistent through the entire trip. Moreover, there are no impossible maneuvers at interchanges/overpasses, meaning that the route is feasible. All 80 routes passed the consistency and feasibility checks.

Insert FIGURE 2 here

Another validation concern was to verify if time-sampling the GPS data deteriorated the accuracy of the derived routes. Recall that, for each trip, on-route GPS data points were sampled at every 5 minute interval (or at the next available time gap). Only those time-sampled GPS coordinates were map-matched. This helped in reducing the computational time needed to map-match each (and every) GPS data point of all trips. To verify if such time-sampling caused any inaccuracies in the derived routes, 45 randomly sampled trips were map-matched without any time-sampling. Next, the same 45 trips were map-matched using time-sampling. Routes for both sets of trips were generated and then compared. It was found that the generated routes for both sets of trips were same in both cases. This suggests that time-sampling the GPS data at a 5-minute interval can reduce the computation time for map-matching while not compromising the accuracy of the derived route. Figure 3 shows an example comparing routes before time-sampling and after time-sampling. As may be observed, the derived route does not change due to time-sampling.

Insert FIGURE 3 here

MEASUREMENT OF TRAVEL ROUTE VARIABILITY

As discussed earlier, the availability of large streams of GPS data offers the opportunity to observe the travel routes for multiple trips between an OD pair. This helps in measuring the extent that the routes taken by different trips overlap with each other (or differ from each other). This section presents a metric to measure the variability (or diversity) of routes chosen by different trips between a given OD pair. The same metric can be used to identify the unique routes used by a large number of trips between an OD pair.

The metric, called *shared link length ratio*, is based on the proportion of the route length shared by other routes, and is computed as below:

$$Shared\ link\ length\ ratio = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} \boldsymbol{l_i^k}}{\sum_{\mathrm{j}=1}^{\mathrm{N}} \boldsymbol{l_j^k}} \qquad (1)$$

where,

$\boldsymbol{l_i^k}$= length of link ***i*** in route ***k*** between an OD pair,
$\boldsymbol{l_j^k}$= length of link ***j*** in route ***k*** between an OD pair,
n = number of links in route ***k*** that are shared with another route for the same OD pair,
N = number of all links in route ***k***.

Using the above-defined *shared link length ratio* metric, the travel routes of all trips between a given OD pair may be categorized into a fewer number of unique routes. In this study, any trip route that shares less than 75 percent of its length with any other route is called a unique route. Specifically, the number of unique routes taken by different trips between a given OD pair is identified as below:

- Step 1: Sort all trips (derived from the data) between the OD pair based on their distance, identify the first route in this order (i.e., minimum length route), and consider it as a unique route.
- Step 2: Get the next trip route and compute its *shared link length ratio* with each (and every) unique route identified so far.
- Step 3: If the *shared link length ratio* of this trip route with all other unique routes is less than 0.75, call the current route unique and add it to the list of unique routes. On the other hand, if the *shared link length ratio* of this trip route with any other unique route is greater than 0.75, the current trip route cannot be called a unique route.
- Step 4: Go to step 2 and repeat until all trips between the OD pair are exhausted.

At the end of the above descried procedure, the travel routes of all trips between a given OD pair are categorized into a smaller number of unique routes taken by those trips. OD pairs with a large (small) number of unique routes may be characterized as OD pairs with high (low) variability in observed routes.

The proposed approach for identifying unique routes was implemented for a sample of 10 different OD pairs, each of which were at least 50 miles apart and had more than 50 trips with derived routes. To better analyze the issue of route variability, trips with detours were not taken into consideration. A detour trip between any given OD pair is typically longer than the majority of routes between the OD pair, perhaps due to intermediate stops that require a detour as opposed to a normal route from the OD pair. After a series of explorations, it was determined that most detour trips between an OD pair tend to be longer than the 95th percentile trip length value of all the trips derived for that OD pair. Therefore, for each OD pair all the trips of length equal to or below the 95th percentile length value were selected to exclude possible detours. Subsequently, unique routes were identified using the approach identified above. Figure 4 shows examples of unique routes identified for four different OD pairs. As may be observed, each OD pair has at least 90 different trips for which the travel routes could be derived (as displayed in the left side panels of the figure). However, the number of unique routes derived for each of the four OD pairs is less than 10. This suggests low variability of observed route choices between the OD pairs. Such low variability could be due to the network structure, where the network provides a limited number of competitive alternative routes between a given OD pair.

Overall, the *shared link length ratio* metric developed in this study is useful for identifying unique routes between a given OD pair. From a practical standpoint, such analysis of route choice variability might shed light on devising route choice set generation algorithms for truck route choice modeling.

Insert FIGURE 4 here

CONCLUSION

Truck probe datasets such as GPS data are being increasingly used for understanding and modeling truck travel patterns at regional, statewide, and megaregional levels. With increasing availability of truck GPS datasets, new opportunities are opening up for analyzing truck route choice behavior. An important precursor step for most analyses and modeling of truck routes using GPS data is to derive the exact travel routes traversed by trucks on the highway network. Any quantitative analysis and modeling of routes requires the measurement of routes in the form of all network links and nodes traversed by the vehicles between the trip origin and destination.

To derive such data, it is important to first map-match the GPS data to the network and then derive the full travel path between the origin and destination. Most existing map-matching approaches that aim at achieving high accuracy are not easy and have not been tested on large, heterogeneous GPS datasets.

This study develops a simple, easy to implement, and effective procedure to derive truck route choice data using truck GPS data. At the heart of this procedure is a map-matching algorithm that is easy to implement and not computationally intensive. The framework is designed to be applicable to a large amount of data with varying frequencies (i.e., varying time gaps) between consecutive GPS records ranging anywhere from as high frequency as a second to as low frequency as 20 minutes. As a large scale demonstration, the framework developed in this research is used to derive routes for over 78,000 truck trips derived from a large GPS dataset (over 100 Million GPS data points) obtained from the American Transportation Research Institute (ATRI). The derived routes are also validated to demonstrate the accuracy of the results.

The study also develops a metric called *shared link ratio length* useful for measuring the amount of overlap between the travel paths of one trip to that of another trip between a given OD pair. In addition, a procedure is developed to employ the metric for measuring the diversity (or dissimilarity or variability) of routes across several trips between a given OD pair and to identify the set of unique routes used between the OD pair. The usefulness of the proposed metric is demonstrated by identifying the unique truck routes used between a few OD pairs.

The route choice data derived in this study can potentially be used for a variety of route choice related analyses in the near future. One such study is to analyze the determinants of the extent of route choice diversity between different OD pairs. From a practical standpoint, the derived data and the route choice diversity analysis could potentially help in devising choice set generation algorithms suitable for truck route choice modeling. Finally, it would be interesting to compare the performance of the proposed map-matching technique with that of other algorithms in the literature.

ACKNOWLEDGEMENTS

The data used for this study was partly derived in a research project funded by the Florida Department of Transportation (FDOT). The opinions, findings, and conclusions expressed in this publication are not necessarily those of FDOT or USDOT. The authors are grateful to ATRI, particularly Jeff Short and Dan Murray from ATRI, for providing access to the raw GPS data. Thanks also to Trang Luong for her assistance in the data analysis.

REFERENCES

1. Strocko, E., Sprung, M., Nguyen, L., Rick, C., & Sedor, J. *Freight Facts and Figures 2013* (No. FHWA-HOP-14-004), 2014.
2. Kim, J. S. Node based map matching algorithm for car navigation system. In *International Symposium on Automotive Technology & Automation (29th: 1996: Florence, Italy). Global deployment of advanced transportation telematics/ITS*, 1996.
3. Ochieng, W. Y., Quddus, M., & Noland, R. B. Map-matching in complex urban road networks. *Revista Brasileira de Cartografia*, *2*(55), 2003.
4. Hess, S., Quddus, M., Rieser-Schüssler, N., & Daly, A. Developing advanced route choice models for heavy goods vehicles using GPS data. *Transportation Research Part E: Logistics and Transportation Review*, *77*, 29-44, 2015.

5. Jagadeesh, G. R., Srikanthan, T., & Zhang, X. D. A map matching method for GPS based real-time vehicle location. *Journal of Navigation*, *57*(03), 429-440, 2004.
6. Pinjari, A.R., Zanjani, A.B., Thakur, A., Nur Irmania, A, Using Truck Fleet Data in Combination with Other Data Sources for Freight Modeling and Planning, BDK84-977-20 Final Report, Florida DOT, 2014.
7. Zanjani, A. B., A. R. Pinjari, M. Kamali, A. Thakur, J. Short, V. Mysore, and S. F. Tabatabaee. Estimation of Statewide Origin–Destination Truck Flows from Large Streams of GPS Data: Application for Florida Statewide Model. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2494,* 2015, pp. 87–96.
8. Thakur, A., Zanjani, A. B., Pinjari, A. R., Short, J., Mysore, V., & Tabatabaee, S. F. Development of Algorithms to Convert Large Streams of Truck GPS Data into Truck Trips. *Transportation Research Record: Journal of the Transportation Research Board,* No. 2529, 2015, pp. 66-73.
9. Yang, J. S., Kang, S. P., & Chon, K. S. The map matching algorithm of GPS data with relatively long polling time intervals. *Journal of the Eastern Asia Society for Transportation Studies*, *6*, 2561-2573, 2005.
10. *Global Positioning System (GPS) Standard Positioning Service Performance Standard* http://www.gps.gov/technical/ps/2008-SPS-performance-standard.pdf. Accessed July 15, 2015.

LIST OF TABLES

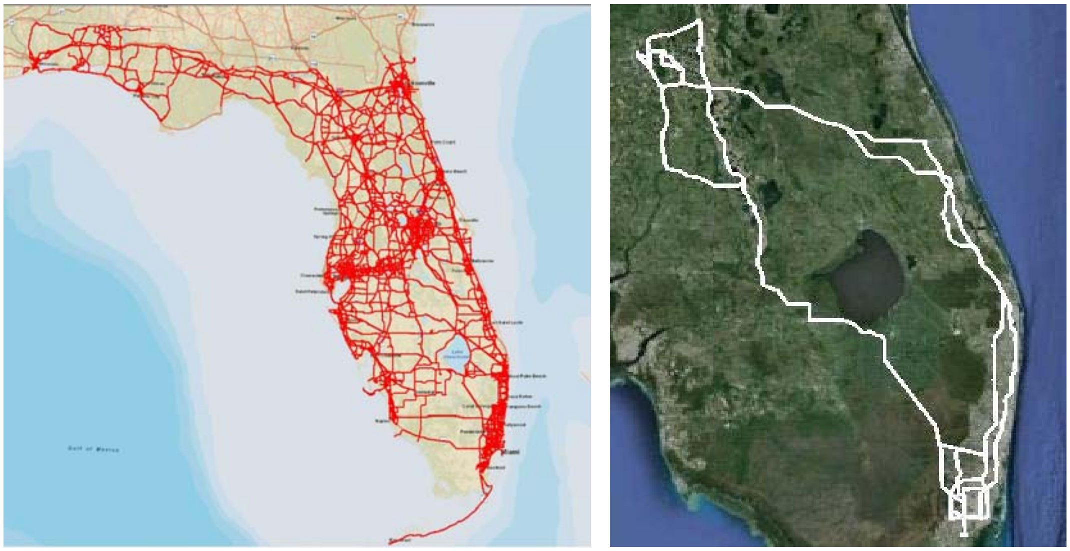
LIST OF FIGURES

TABLE 1 Cross-Tabulation between Largest Time Gap and Its Corresponding Spatial Distance for GPS Data with Spot Speed

Spatial gap (Miles) Time gap (Minutes)	< 1	1-5	5-15	15-20	20-25	25-30	30 <	Sum
< 1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
1-2	0.3%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.5%
2-5	2.7%	14.2%	0.3%	0.0%	0.0%	0.0%	0.0%	17.2%
5-14	4.3%	11.2%	3.6%	0.0%	0.0%	0.0%	0.0%	19.2%
14-15	0.8%	4.1%	24.5%	3.9%	0.0%	0.0%	0.0%	33.2%
15-20	4.5%	0.7%	1.6%	0.1%	0.0%	0.0%	0.0%	6.8%
20-25	3.4%	0.1%	0.3%	0.1%	0.0%	0.1%	0.0%	4.0%
25-30	2.4%	0.2%	0.4%	0.1%	0.1%	0.1%	0.1%	3.4%
30-45	2.1%	0.1%	0.2%	0.0%	0.1%	0.1%	0.1%	2.6%
45 <	12.0%	0.2%	0.3%	0.0%	0.0%	0.0%	0.5%	13.0%
Sum	32.6%	31.0%	31.0%	4.2%	0.2%	0.2%	0.6%	100.0%

TABLE 2 Cross-Tabulation between Largest Time Gap and Its Corresponding Spatial Distance for GPS Data without Spot Speed

Spatial gap (Miles) Time gap (Minutes)	< 1	1-5	5-15	15-20	20-25	25-30	30 <	Sum
< 1	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
1-2	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%
2-5	0.2%	0.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.6%
5-14	0.3%	3.3%	0.6%	0.0%	0.0%	0.0%	0.0%	4.2%
14-15	0.0%	0.4%	0.2%	0.0%	0.0%	0.0%	0.0%	0.6%
15-20	0.1%	1.4%	1.7%	0.2%	0.0%	0.0%	0.0%	3.5%
20-25	0.1%	0.8%	1.8%	0.2%	0.1%	0.0%	0.0%	3.1%
25-30	0.1%	0.5%	1.9%	0.5%	0.3%	0.2%	0.1%	3.6%
30-45	0.3%	0.6%	3.2%	1.8%	1.5%	1.3%	1.4%	10.1%
45 <	18.1%	1.0%	3.7%	1.9%	2.4%	3.3%	44.0%	74.3%
Sum	19.2%	8.3%	13.2%	4.7%	4.3%	4.7%	45.5%	100.0%



(a) Routes generated for all 78,381 trips (b) Routes generated between one O/D pair

FIGURE 1 Routes derived in the study

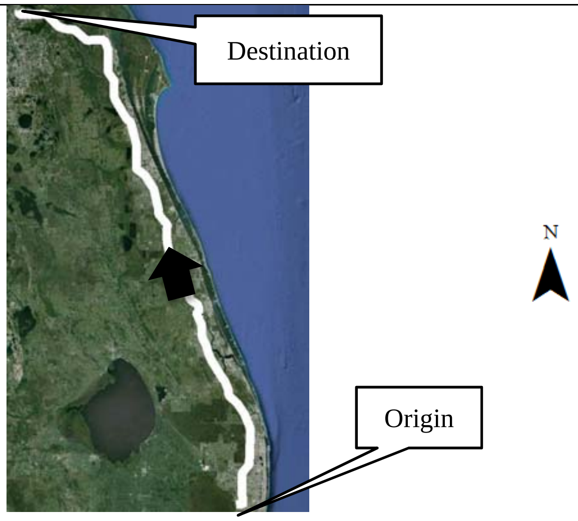
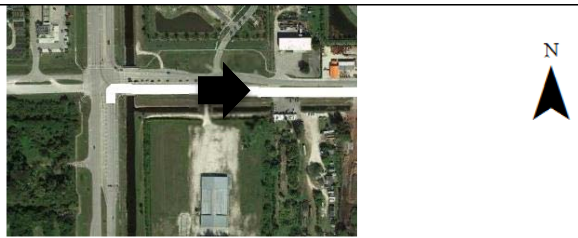
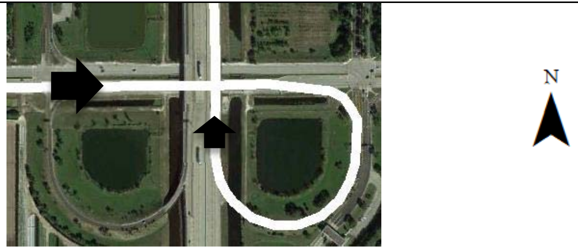
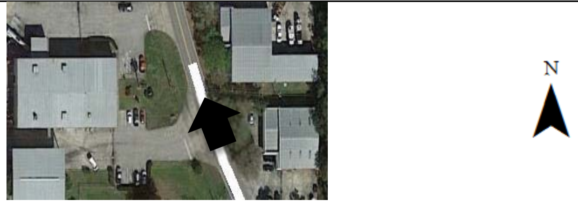
Location	
Route	
Origin	
Interchange/overpass	
Destination	

FIGURE 2 Consistency and feasibility checks for a derived trip route

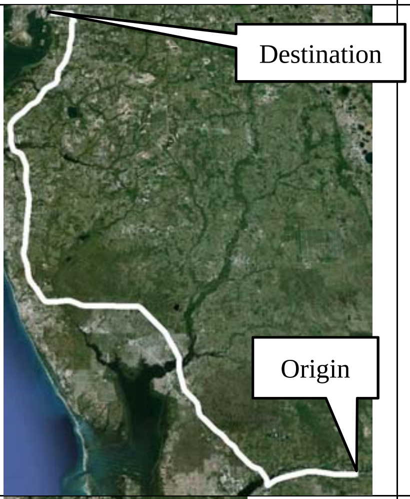
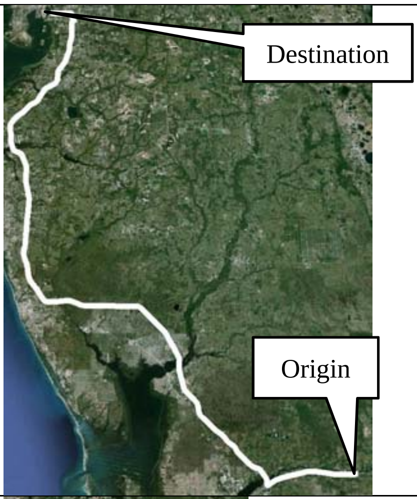
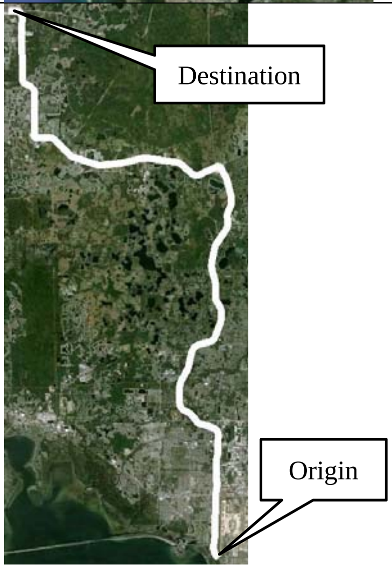
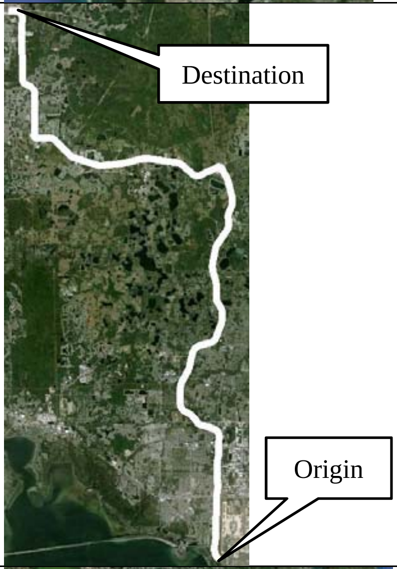
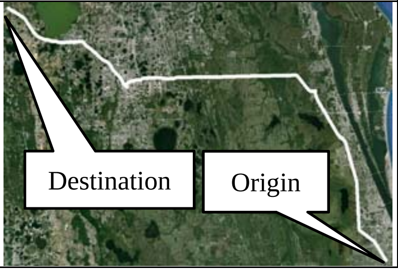
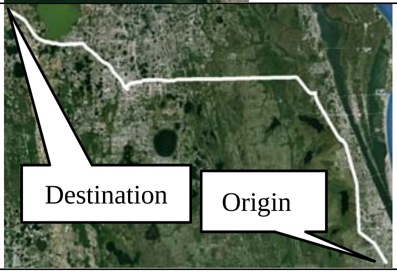
Case number	Before time-sampling	After time-sampling	Are the routes similar?
1			Yes
2			Yes
3			Yes

FIGURE 3 Routes with and without time sampling

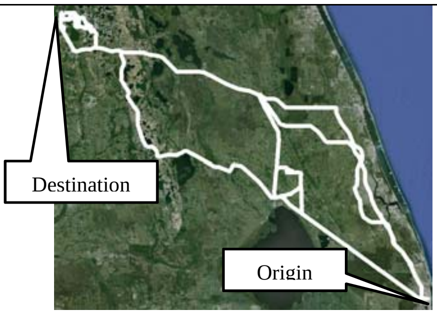
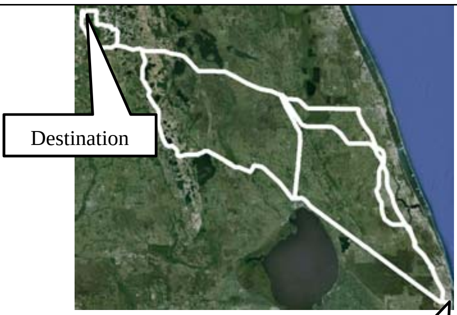
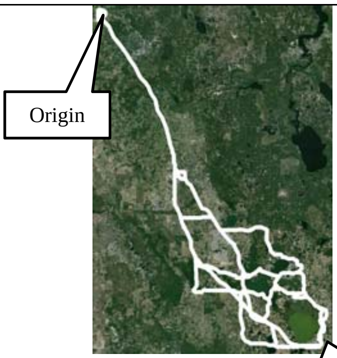
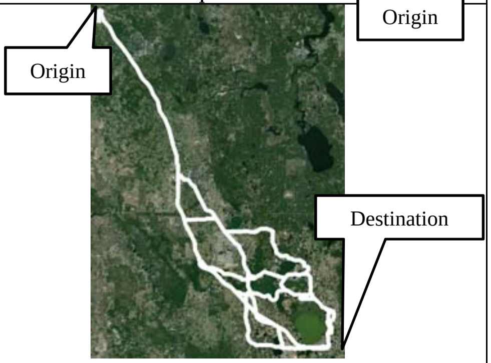
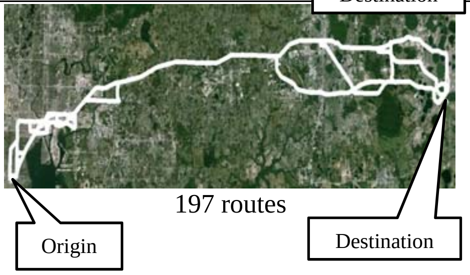
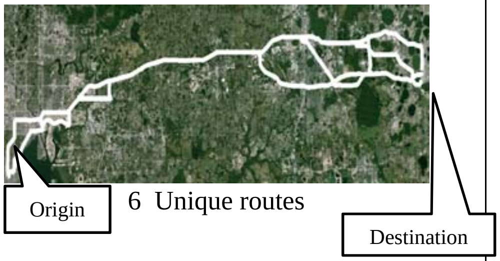
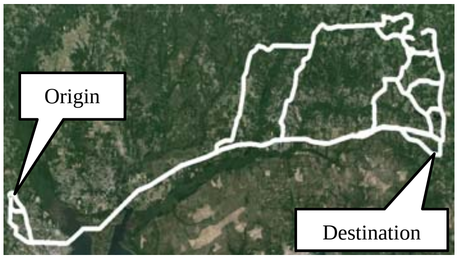
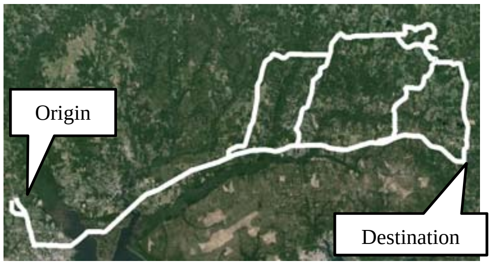
Case number	Routes of all trips of length 95 percentile value or less	Unique routes derived for the OD pair
Case 1: Minimum trip distance = 156 miles	 91 routes	 4 Unique routes
Case 2: Minimum trip distance = 117 miles	 237 routes	 10 Unique routes
Case 3: Minimum trip distance = 55 miles	 197 routes	 6 Unique routes
Case 4: Minimum trip distance = 91 miles	 94 routes	 6 Unique routes

FIGURE 4 Unique routes derived for a sample of 4 OD pairs