Commercial Vehicle Travel Time Estimation in Urban Networks using GPS Data from Multiple Sources

Ender Faruk Morgul, M.Sc. (Corresponding Author)
Graduate Research Assistant
Rutgers Intelligent Transportation Systems (RITS) Laboratory
Dept. of Civil and Environmental Engineering
Rutgers, The State University of New Jersey
96 Frelinghuysen Road, Piscataway, NJ 08854
Tel: (732) 445-5496, Fax: (732) 445-0577
Email: morgul@eden.rutgers.edu

Kaan Ozbay, Ph.D
Professor, Dept. of Civil and Environmental Engineering
Rutgers Intelligent Transportation Systems (RITS) Laboratory
Rutgers, The State University of New Jersey
96 Frelinghuysen Road, Piscataway, NJ 08854
Tel: (732) 445-2792 Fax: (732) 445-0577
Email: kaan@rci.rutgers.edu

Shrisan Iyer, M.Sc.
Research Engineer
Rutgers Intelligent Transportation Systems (RITS) Laboratory
Rutgers, The State University of New Jersey
96 Frelinghuysen Road, Piscataway, NJ 08854
Tel: (732) 445-6243 x274
Email: shrisan@rci.rutgers.edu

José Holguín-Veras, Ph.D., P.E.
Professor, Dept. of Civil and Environmental Engineering
Rensselaer Polytechnic Institute
110 Eight St., Troy, NY 12180
Tel: (518) 276-6221 Fax: (518) 276-4833
Email: jhv@rpi.edu

Word count: 4,611 + 4 Tables + 5 Figures = 6,861 words
Abstract: 220 words

Paper Submitted for Presentation and Publication at the
Transportation Research Board’s 92nd Annual Meeting, Washington, D.C., 2013
Re-submission Date: November 15th, 2012
ABSTRACT

Realistic travel time estimation for urban commercial vehicle movements is challenging due to limited observed data, large number of Origin-Destination (OD) pairs, and high variability of travel times due to congestion. Moreover most traditional data collection methods can only provide information in an aggregated form which is not sufficient for micro-level analysis. On the other hand, the usage of Global Positioning Systems (GPS) data for traffic monitoring and planning has been continuously growing with significant technological advances in the last two decades.

In this paper we provide a comprehensive review of the current usage of GPS data in transportation planning applications and present a practical integrated methodology for using a robust source of GPS data, for commercial vehicle travel time prediction. A comparison with observed truck travel times collected from a limited source of truck-GPS data reveals that travel times obtained from taxi-GPS data approximate those of trucks, and can be used to supplement truck-GPS travel time data on a wider scale. While the amount of truck-GPS data is limited to a small number of trucks serving very few OD pairs, taxi-GPS provide citywide penetration and can estimate travel times between most OD pairs in a city. The provided methodology leads to simple and effective travel time estimations using taxi-GPS data without a need for an extra data collection effort.
INTRODUCTION

Accurate estimation of in-city commercial vehicle movements is an essential part of transportation planning in urban areas. Time-of-day dependent travel time estimation is not only an important issue that needs to be taken account by transportation decision-makers but also a critical element for carriers’ strategic plans such as fleet sizing and dispatching. In particular, being a part of a highly competitive industry, urban delivery companies must consider routes with short and reliable travel times (which may differ by time of day) to improve service quality and customer satisfaction.

Travel times in urban areas are significantly subject to variations throughout the day, and specific locations may suffer from recurrent congestion. For the freight industry, poorly designed routing estimations that ignore travel time variations direct delivery trucks to congested arterials and streets, which in turn causes extra costs to companies related to labor and customer delays. Furthermore, slow moving heavy vehicles in urban traffic aggravate traffic conditions for all users, and increase negative externalities such as emissions and noise. Supply chain strategies cannot be based on distances between customers. Distance alone is not a good enough predictor for urban travel time estimation due to the differences in infrastructure and connectivity characteristics (e.g. turning movements) of the links. Therefore generalizing limited observed data to the entire network based on trip distances is almost impossible (1).

Understanding movements of commercial vehicles in urban areas is not an easy task because of the limitations of collecting a sufficient amount of disaggregate data (2). Despite the fact that the importance of precise travel time information is well addressed for both movements of people and freight, traditional traffic surveillance methods such as loop-detectors or automatic vehicle identification systems are not able to deliver comprehensive data for network-wide analysis. Moreover, fixed-location equipment used for these methods come at high cost which makes them infeasible to install and maintain in places other than major corridors (1). Emerging technologies such as Global Positioning Systems (GPS) on the other hand, offer promising improvements in traffic monitoring technology and have become widely available within the last decade. Probe vehicles equipped with relatively cheap devices provide researchers with vast amounts of disaggregate traffic data that can be easily used for determining link travel speeds and detecting congested locations and traffic delays. Implementation of GPS-based tracking of taxi and bus systems has been growing in major cities around the world and GPS-based probe vehicles have been used as traffic sensors for real-time and historical data in many different transportation problems; including taxi dispatching (3, 4), cost-benefit analysis for snow operations (5), and evaluation of road safety projects (6). Similarly, over the last two decades many freight companies have replaced their traditional sheet-based trip logs by GPS recordings which provide disaggregate information about trip chains, trip and customer service durations, distances, and routes (7).

Although the freight industry is aware of the importance of GPS-based vehicle tracking for commercial purposes, companies cannot or will not provide their information for planning or research purposes. Customer privacy and strategic concerns prevent companies from sharing their GPS data and therefore available data is limited to a small set of specific regions. However where industry or mode-specific data is lacking, other sources of GPS data may be available to supplement available datasets. In the case of urban transportation, shared roadways and often-
congested conditions lead to a somewhat homogenous distribution of traffic speeds and travel times across vehicles.

This research is motivated by the importance of determining network-wide time-dependent travel times for urban commercial vehicles, and attempts to use GPS data collected from taxis to supplement the limited amount of available freight transportation data. We present an empirical methodology for screening the noisy raw GPS data obtained from two different sources for the same network and develop time-based clusters to compare the travel times between origin-destination pairs for specific time intervals. Our approach leads to accurate commercial vehicle travel time estimations for regions where observed commercial vehicle data is limited or do not exist, by using taxi travel times that cover nearly the entire network. Realistic travel time estimations can be used in the calibration process for simulation-based studies and as a result enhance the validity of travel times.

LITERATURE REVIEW

A significant amount of research has been conducted using GPS-equipped vehicle-based data, or floating car data, as referred to in some of the literature. The majority of these studies focus on routing optimization and traffic performance analysis, such as real-time travel time prediction or link speed estimations. Some studies have also been conducted for transportation planning purposes. A summary of relevant literature is given in Table 1 along with the sample sizes they used. In this section we review the existing literature in two categories: 1) studies utilizing GPS-probe vehicle data for traffic monitoring 2) transportation planning studies using GPS data.

Traffic Monitoring

Probe vehicles have been used for traffic monitoring for several decades, however the portion of these vehicles in traffic made the adequacy of the data questionable (8). Several studies have been conducted to determine the optimal number of probe vehicles needed to reflect realistic traffic conditions, but the growing implementation of GPS tracking of taxis, buses, and other vehicles has added an enormous amount of equipped vehicles traveling through the urban traffic networks. Sanwal and Warland (1995) underlined the insufficient amount of sensors available for traffic surveillance and proposed idea of using vehicles as sensors. The authors gave algorithms for travel speed calculation and their simulation results showed that probe vehicles accounting for approximately 4% portion of total traffic are required for satisfactory estimations (9).

The following studies tried to explore the opportunities provided by GPS-based probe vehicle data. Liao (2001) conducted a survey for taxi drivers using GPS devices for dispatching in Singapore and reported that there was a significant improvement in service speed and operation accuracy compared to traditional methods for taxi dispatching that are highly dependent on operators’ experience (4). Schafer et al. (2002) reviewed one of the first examples of taxi-GPS implementations in Europe and concluded that real-time traffic information obtained taxis has a high potential for further usage (10). Lorkowski et al. (2005) considered taxi-GPS data processing steps such as data filtering and map matching and discussed potential applications such as dynamic routing and automatic congestion detection. Following his empirical analysis using taxi-GPS data from 700 vehicles in Stuttgart, Germany, the author reported that probe vehicle amount about 1% of total traffic is reasonable to estimate traffic conditions (11).
### TABLE 1 Literature using GPS-Based Probe Vehicle Data

<table>
<thead>
<tr>
<th>Authors</th>
<th>Year</th>
<th>Data Size</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Traffic Monitoring</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zou et al. (12)</td>
<td>2005</td>
<td>100 taxis</td>
</tr>
<tr>
<td>Brockfield et al. (19)</td>
<td>2007</td>
<td>700 taxis</td>
</tr>
<tr>
<td>Lahrmann (24)</td>
<td>2007</td>
<td>1 month taxi-GPS data, 10 taxis</td>
</tr>
<tr>
<td>Liu et al. (14)</td>
<td>2008</td>
<td>4000 taxis</td>
</tr>
<tr>
<td>Sananmongkhonchai et al. (15)</td>
<td>2008</td>
<td>1 month taxi-GPS data, 4000 taxis</td>
</tr>
<tr>
<td>Hunter et al. (16)</td>
<td>2009</td>
<td>2 months taxi-GPS data, 50 taxis</td>
</tr>
<tr>
<td>Li et al. (27)</td>
<td>2009</td>
<td>57 days taxi-GPS data, 7000 taxis</td>
</tr>
<tr>
<td>Liu et al. (21)</td>
<td>2009</td>
<td>6 months taxi-GPS data, 249 taxis</td>
</tr>
<tr>
<td>Herring et al. (17)</td>
<td>2010</td>
<td>3 months taxi-GPS data, 500 taxis</td>
</tr>
<tr>
<td>Yuan et al. (18)</td>
<td>2010</td>
<td>3 months taxi-GPS data, 33000 taxis</td>
</tr>
<tr>
<td>Ehmke et al. (1)</td>
<td>2012</td>
<td>2 years taxi-GPS data</td>
</tr>
<tr>
<td>Miwa et al. (25)</td>
<td>2012</td>
<td>6 months taxi-GPS data, 1500 taxis</td>
</tr>
<tr>
<td>Yokota and Tamagawa (26)</td>
<td>2012</td>
<td>1 month truck-GPS, 300 trucks</td>
</tr>
<tr>
<td><strong>Transportation Planning</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Li et al. (22)</td>
<td>2009</td>
<td>57 days taxi-GPS data, 7000 taxis</td>
</tr>
<tr>
<td>Uno et al. (23)</td>
<td>2009</td>
<td>10 day bus-GPS data, 100 buses</td>
</tr>
<tr>
<td>Kinuta et al. (6)</td>
<td>2010</td>
<td>6 months taxi-GPS, 1 month truck-GPS, 70 taxis, 130 trucks</td>
</tr>
<tr>
<td>Munehiro et al. (5)</td>
<td>2012</td>
<td>1 year taxi-GPS data, 115 taxis</td>
</tr>
</tbody>
</table>

A primary area of research that employs taxi-GPS data is the estimation of link-based travel speeds on arterials. Zou et al. (2005) analyzed taxi-GPS data from 100 vehicles in Guangzhou, China and provided a methodology for arterial speed estimation, including data screening for improving data quality by removing irrelevant data points. In the study when the number of probe vehicles accounted for 3% of total traffic they gave significantly lower errors in travel speed estimations (12). Lee et al. (2006) used the Fuzzy c mean method for clustering taxi-GPS data and estimated the arterial speeds. They concluded that even a small amount of data (10 days) can be a good estimator for actual speed distributions (13). Liu et al. (2008) developed a tool for real-time traffic information based on link speeds using taxi-GPS data (14). Sananmongkhonchai et al. (2008) considered the combination of real-time taxi data with historical hourly speed profiles and observed improvements in travel speed estimations (15). Network-wide traffic analyses were conducted by several authors utilizing taxi-GPS data. Hunter et al. (2009) used 2 months taxi data from San Francisco, CA for travel time and trip routing estimation (15). In particular, they considered the parts of the network that are not used frequently and suffer from a limited amount of taxi-GPS data. The authors tried to predict the most likely trajectories using a learning algorithm and tested their algorithm using taxi-GPS data. The results showed that the developed algorithm, which assumed every link has an independent travel time distribution, successfully estimated morning and afternoon peak traffic congestion. Herring et al. (2010) estimated traffic conditions using taxi-GPS data with a relaxed assumption that travel times are uniformly distributed over a link. Results of the analysis showed that the accuracy improved by 36.9% over a baseline approach in which average travel speeds are...
assigned to the links for each observation (17). Yuan et al. (2010) used historical taxi-GPS data for determining the fastest route for navigation purposes. Taxi travel times were clustered for different time slots and travel time distributions were determined. Approximately 70% of the routes detected were found to be faster than the routes found by other methods (18). Ehmke (2012) provided basic steps for long period taxi-GPS data analysis focusing on travel speed and travel time detection for large networks. The methodology included two-staged clustering first by time-of-day and second by link speed variations. (1).

The validity of data collected from GPS-based probe vehicles was also analyzed by several researchers. Brockfield (2007) investigated the validity of computed travel times from taxi-GPS data with test drives and found that 80% of compared travel times were within less than 10% error, equivalent to 2 minute deviations for an average travel time of 20 minutes (19). Poomrittigul et al. (2008) studied the correlation between mean travel speed, which required automatic vehicle matching using fixed sensors, with time mean speed, which was obtained from GPS data. They provided a grouping method for the GPS data to reduce the variance in travel speeds and then calculated the time mean speed. The correlation of data for the two methods was found to be 0.94, which showed that using taxis as traffic sensors can also be effective (20). Liu et al. (2009) conducted a feasibility analysis for traffic information data collection using GPS-equipped vehicles. The authors concluded that measurement errors arising from GPS equipment itself had a small impact on traffic monitoring analysis, whereas communication errors (e.g. lost signal, multiple recording etc.) needed to be filtered carefully before data analysis. They also stated that the taxi-GPS data used was not enough for real-time traffic information gathering, especially for link travel times, but historical data was usable to detect traffic congestion efficiently (21).

Transportation Planning

GPS data has not been utilized in transportation planning studies as effectively as it has been in studies focusing on traffic performance measurement. Li et al. (2009b) gave a methodology for determination of recurrent congestion points using historical taxi-GPS data and validated the results with real-time traffic measurements (22). Uno et al. (2009) used data from 100 buses in Hirakita, Japan for estimating travel time variability and level of service (LOS) of roads. Deceleration/acceleration before/after bus stops were taken into account in their model and the increase in travel time due to stopping was eliminated while carrying out the estimations (23). Kinuta et al. (2010) discussed the usefulness of probe car data in evaluation of road-safety project impacts (6). Munehiro et al. (2012) conducted a cost-benefit analysis for snow and ice operations using travel speeds from 1 year of taxi-GPS data and showed that benefits exceeded the cost of these operations (5).

This study aims to provide insight into understanding commercial vehicle movements by finding a relation between truck movements and taxi movements within the same network. Currently, available truck-GPS data is still not sufficient to investigate the majority of freight movements in a city since most companies are reluctant to disclose this kind of sensitive information for commercial reasons, or simply do not collect and store such data. Therefore an observed relation between truck and taxi movements will lead to reliable estimation of travel times for commercial vehicles for the regions where truck data is not present. The following section explains the methodology for using taxi-GPS data to achieve this goal.
RESEARCH METHODOLOGY

Analysis of commercial vehicle movements in urban networks generally requires consideration of a very large number of dense OD pairs compared to inter-city trips. However, existing data is not sufficient enough for network-wide analysis due to the lower number of commercial vehicles with respect to all other vehicles, and observations only represent a limited portion of actual trips. Figure 1 shows the observed data for truck delivery stop locations obtained from a major delivery company serving businesses in Manhattan, New York City. It can be seen that majority of the points are in southern Manhattan while there are only a few zones with data in northern parts which leaves 67% of the network unexplored.

In this section, we provide a practical methodology for estimation of truck travel times using taxi-GPS data, and provide a statistical comparison of available truck trip times with taxi trip times for the same OD pairs. The null hypothesis that is tested in this paper is $H_0$: the average OD truck trip time based on observed GPS data for a given OD pair “ij” is not statistically different than the average trip travel times estimated using taxi-GPS data. With an alternative hypothesis of $H_1$: the average truck trip travel times obtained from truck-GPS data is statistically different from average trip travel times obtained from taxi-GPS data.

Failure to reject $H_0$ indicates that taxi-GPS data is a viable option for estimating commercial vehicle travel times for regions where truck-GPS data is not available. Alternatively, travel time distributions can also be used depending on the size of the data. In this case, a significant number of OD pairs with available truck-GPS data to conduct a distribution analysis is not available, therefore average values are employed. For taxi-GPS data, a full set of New York City taxi-GPS data is available. For this paper, one month of taxi-GPS data consisting of approximately 3.5 million records is used, which is comparable to the earlier studies shown in Table 1. Statistical comparisons for hypothesis testing are carried out with average taxi travel times and travel time distributions are used for further analysis.
FIGURE 2 Data Analysis Flowchart.

Figure 2 shows the steps followed in data analysis, which begins with the determination of customer stops from truck-GPS data which constitute the origin/destination points for travel time comparisons. Stop points are then clustered by location and travel times are calculated between clusters. Next, taxi-GPS data are processed to find the trip start/stop points by previously determined clusters and taxi travel time distributions are determined. Then the correlation between the two sets of travel times are analyzed by statistical methods.

Data

Truck-GPS recordings from a major delivery company were collected for a total of 6 months in 2009 and 2010 by monitoring 5 delivery trucks serving the New York metropolitan area. Location information along with different fields such as operation status as the time of arrival to a customer (“stop”), departure (“start”) and ignition on/off status are recorded with 2 minutes intervals throughout the delivery tour. Data collection is carried out in two ways: 1) driver activation is required by a simple button tap for indicating delivery tour start, stop or moving conditions and 2) passive GPS recording for routing after driver pushes the tour start button. This method is advantageous since once the recording is completed, data analysis is much easier than with a continuous passive recording, since customer stops are clearly marked by drivers’ manual effort. One of the earlier studies that used passive GPS data came up with a trip-identification algorithm that tried to identify customer stops by checking GPS coordinates (7). However this

TRB 2013 Annual Meeting Paper revised from original submittal.
type of analysis may not be applicable to larger datasets because of the risk of identifying false
customer stops when the truck is not moving due to congestion or traffic signals. Therefore in
our methodology, manual input of drivers provides extra proof for trip identification which
enhances the reliability of the data analysis.

Taxi-GPS data provided by NYC Taxi and Limousine Commission (TLC) includes more than 80
million records integrated to a single database. A trip is defined as the time period from when a
customer enters and leaves the taxi, therefore empty trips are not included. Data are recorded per
trip and include fields such as trip start/stop times, trip duration, location of origins/destinations,
and trip fare. Routing information is not available since it is not a continuous passive recording.
In this study we use 1 month data from January 2009, which covers more than 13,000 taxis
covering all time periods and almost all regions in New York City. According to the calibrated
New York Best Practice (NYBPM) travel demand model, taxi traffic accounts for 11.9% of total
traffic flow in Manhattan (28).

Clustering
GPS data points need to be aggregated for useful analysis. Two types of clustering are
performed; first by time-of-day and next by location. Aggregation level should be chosen
carefully since too fine clustering of data points may lead to insufficient amount of data per
cluster whereas larger clusters may lead difficulties in interpreting the results due to possible loss
of detailed cluster-specific characteristics. Time-of-day clustering is done according to four time
periods (AM: 6am-10am, Midday: 10am-3pm, PM: 3pm-7pm, Night, 7pm-6am). Readily
available transportation analysis zones (TAZs) from the NYBPM provide spatial data
aggregation for transportation planning studies (28). Location based clustering is performed by
simple GIS geocoding techniques. A total of 318 zones are used for Manhattan which covers a
total of 33.8 sq mi. One zone is approximately 0.1 sq. mi. large, and covers five city blocks on
average. Figure 3 overlays the cluster zones and the road network in the study area.

Data Processing
A trip identification algorithm is generated for truck-GPS data to determine the exact trip
start/stop locations and times and filter trips outside of Manhattan. The algorithm checks the
manual entries for vehicle statuses along with GPS coordinate information to ensure that the
truck made a stop for a delivery. A total of 1,782 trips are identified from the truck-GPS dataset
and these trips are classified by time of day. A total of 331 Origin-Destination pairs are found.

Taxi-GPS is clustered by the Origin-Destination pairs that are found from truck-GPS data and
result in a total of 7,234 trips for the 331 origin-destination pairs. Table 2 shows the number of
matching OD pairs by time period and total number of trips found from both datasets. 99% of the
OD pairs found based on truck customer stops could be matched in the taxi data for the same
time periods. A total of 331 OD pairs are found which covers 33% of the entire network.
FIGURE 3 Manhattan Zones (28).

TABLE 2 Sample Size

<table>
<thead>
<tr>
<th>Time Period</th>
<th>Matching OD Pairs</th>
<th>Number of Trips (Truck)</th>
<th>Number of Trips (Taxi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM (6am-10am)</td>
<td>117</td>
<td>724</td>
<td>1667</td>
</tr>
<tr>
<td>Midday (10am-3pm)</td>
<td>72</td>
<td>247</td>
<td>1423</td>
</tr>
<tr>
<td>PM (3pm-7pm)</td>
<td>4</td>
<td>9</td>
<td>38</td>
</tr>
<tr>
<td>Night (7pm-6am)</td>
<td>138</td>
<td>802</td>
<td>4106</td>
</tr>
<tr>
<td>TOTAL</td>
<td>331</td>
<td>1782</td>
<td>7234</td>
</tr>
</tbody>
</table>
Filtering

Data rejection is an important step for reliable GPS data analysis, as pointed out in many of the earlier studies (1, 7, 29). Major sources for false travel time measurements from GPS readings detected in this study are:

- Driver error / False entry
- GPS device malfunctioning
- Unrealistically short trips between zones (e.g. Truck changing parking spot, delivery delays, taxi trips shorter than 1 min.)
- Non-recurring traffic conditions (e.g. accidents, road work)

False measurements are generally observed as outliers in a travel time distribution between an origin-destination pair. One of the most widely used methods for outlier detection, Chauvanet’s criterion, is applied to all calculated travel time distributions obtained after data processing. Chauvanet’s criterion assumes a normal distribution of the data and calculates the probability of occurrence of every single measurement in the distribution (30). If there are N measurements $X_1, X_2, ..., X_N$ and $X_{sus}$ is the suspicious measurement to be tested, first the deviance from the mean is calculated using

$$t_{sus} = \frac{|X_{sus} - X_{mean}|}{SD_X}$$  \hspace{1cm} (1)

Where $X_{mean}$ is the mean and $SD_X$ is the standard deviation of all measurements. Then the probability of getting a measurement as deviant as $X_{sus}$ is calculated by

$$n = N \times Pr(outside t_{sus}, SD)$$ \hspace{1cm} (2)

In equation 2, $n$ is called the threshold value, which is considered as 0.5 in most applications. Therefore according to Chauvanet’s criterion, if $n < 0.5$, the measurement $X_{sus}$ can be rejected. This method is found to be simple and effective for automatic detection of most outliers. However manual detection is also useful for distributions that have a low number of measurements. For example, in our case it was observed that within-cluster measurements show significant variation for both datasets. Some of the detected reasons include delivery trucks searching for a parking spot or making multiple stops for the same customer, resulting in very low travel times, or, taxis having extremely long trips within the same cluster due to driver error, device malfunctioning, or a change in destination. Therefore in-cluster trips are excluded from the travel time analysis and comparisons are done for trips between different clusters.

RESULTS

Travel times for each trip are calculated and the corresponding zones are considered as origin-destination pairs for both datasets. Hypothesis testing for statistical comparison of median travel time from both taxi and truck data is carried out by using the Wilcoxon signed-rank test for each OD pair. Table 3 shows the number of OD pairs and the percentages to the total sample for which the null hypothesis can be rejected at the 95% confidence interval. According to the analysis, the null hypothesis cannot be rejected at the 95% confidence level for 60% of the observed OD trips in AM period. For the Midday period, the null hypothesis cannot be rejected at the 95% confidence level for 39% of the trips. PM period trips are the lowest in sample size
due to the lower frequency in truck delivery operations and the null hypothesis cannot be rejected for 50% of the available data. Night period trips exhibit the lowest rejection rate, as the null hypothesis could not be rejected at the 95% confidence level for only 29% of the trips. These results show that daytime trip travel times associated with commercial deliveries, particularly those in the AM and PM Peak periods, can be better estimated by taxi-GPS data compared to the night period. This is most likely due to the nature of night time traffic conditions, where the network is closer to free flow conditions and vehicle characteristics or driver behavior is more likely to affect travel speeds. Since trucks generally move slower than automobiles it is likely that they would experience slower speeds and higher travel durations than taxis during free flow conditions. This is also exhibited to a lesser degree in the Midday period.

**TABLE 3 Wilcoxon Signed-Rank Test Results**

<table>
<thead>
<tr>
<th>Number of OD Pairs that H₀ is not rejected</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>Midday</td>
</tr>
<tr>
<td>70</td>
<td>28</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Percent of OD Pairs that H₀ is not rejected</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>Midday</td>
</tr>
<tr>
<td>60%</td>
<td>39%</td>
</tr>
</tbody>
</table>

Hypothesis testing is based on strong assumptions about the distribution of travel times and assumed critical values can be too strict to reject the null hypothesis. Therefore further comparison of travel times is conducted between the dispersion of travel time distributions for taxi-GPS data and central tendency of travel times for truck-GPS data. Measures of dispersion can be conducted in different ways such as standard deviation or inter-quantile difference (e.g. 90th-50th or 75th-50th). Measures of central tendency, on the other hand, include the mean or the median of travel times. It is assumed that for comparison purposes the upper tail of taxi travel time distributions is more meaningful for in-city traffic since travel speeds falling in the lower tail generally represent free flow conditions. Using this data may lead to an underestimation for realistic travel times. The probability of being exposed to delays is higher in city traffic due to several disturbances due to traffic signals or congestion, therefore the 90th-50th percentile range is chosen as the dispersion criteria. For truck-GPS data mean travel times are considered due to the smaller size of the dataset. Thus the criteria for comparison is developed as: if the mean truck travel time falls in an envelope between the 90th percentile and 50th percentile of the taxi-GPS travel time distributions, it is assumed to be a good estimation. When truck travel times do not fall into this range errors are calculated based on the difference between the closest strip of the envelope, that is either the 90th percentile value when the truck travel time is higher or the 50th percentile value when the truck travel time is lower.
FIGURE 4 Travel Time Comparisons.
Figure 4 shows the comparison results for the AM, Midday, and Night period results (PM Peak is eliminated due to low sample size). Error analysis shows that 45% of truck travel times for AM period are in the 90th-50th percentile envelope of taxi travel time distribution, and 50% and 33% for the Midday and Night periods respectively. Higher accuracy is observed in the daytime periods compared to the Night period. Table 4 summarizes the results and it is seen that a significant portion of OD travel times can be estimated within 1 minute of error. Moreover, almost all of the observed travel times are estimated within 5 minute error. Keeping in mind that these errors are strongly dependent on the observed truck trip duration, the results are very promising if a 1 minute or even 5 minute difference can be an acceptable variance for in-city trips.

**TABLE 4 Comparison Results-Truck OD Travel Times vs Taxi OD Travel Times**

<table>
<thead>
<tr>
<th></th>
<th>Number of OD Pairs</th>
<th>Inside Envelope (50th-90th percentile)</th>
<th>&lt;1 min. error</th>
<th>&lt;2 min. error</th>
<th>&lt;5 min. error</th>
<th>&gt;5 min. error</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TOTAL</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AM (6am-10 am)</td>
<td>117</td>
<td>53</td>
<td>62</td>
<td>92</td>
<td>117</td>
<td>-</td>
</tr>
<tr>
<td>MD (10am-3pm)</td>
<td>72</td>
<td>36</td>
<td>43</td>
<td>53</td>
<td>65</td>
<td>7</td>
</tr>
<tr>
<td>PM (3pm-7pm)</td>
<td>4</td>
<td>-</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>-</td>
</tr>
<tr>
<td>NT (7pm-6am)</td>
<td>138</td>
<td>46</td>
<td>70</td>
<td>102</td>
<td>136</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Inside Envelope (50th-90th percentile)</th>
<th>&lt;1 min. error</th>
<th>&lt;2 min. error</th>
<th>&lt;5 min. error</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>45%</td>
<td>53%</td>
<td>79%</td>
<td>99%</td>
</tr>
<tr>
<td>MD</td>
<td>50%</td>
<td>60%</td>
<td>74%</td>
<td>90%</td>
</tr>
<tr>
<td>PM</td>
<td>-</td>
<td>25%</td>
<td>50%</td>
<td>100%</td>
</tr>
<tr>
<td>NT</td>
<td>33%</td>
<td>51%</td>
<td>74%</td>
<td>99%</td>
</tr>
</tbody>
</table>
FIGURE 5 Distribution by Travel Time for Time Periods.
Figure 5 shows the error for each OD pair and their actual travel times from the truck-GPS data. The trips are sorted by ascending truck travel times and it is observed that for daytime periods (e.g. AM and Midday) higher errors are concentrated mainly in the two ends of the graph corresponding to the lowest and highest travel times. Mid-range travel time values are generally estimated better considering the smaller error percentages to the actual truck measurements.

Night period errors are distributed more evenly compared to daytime and it can be observed that even trips that are relatively longer can be estimated with small error using taxi travel times. However, it should be noted that in our case the exact form of the error distributions is difficult to obtain by trip duration and time of day.

Comparison results show that the developed methodology can be useful for estimating commercial vehicle travel times using the extensive amount of taxi-GPS data available. A more detailed analysis for error distributions, possibly using more data points from observed truck movements, enables generalization of the results for the remaining regions of the network. This way of analysis leads to many different applications for both delivery industry and transportation decision-makers.

CONCLUSION

As addressed by several studies, data associated with commercial vehicle movements are sparse and not always easily reachable (1, 7). The idea of employing taxis as probe vehicles for transportation studies is becoming increasingly popular and has been implemented for various purposes such as real-time routing, link speed estimation, and travel time estimation. This paper provides a practical methodology for time-dependent commercial vehicle travel time estimation for planning purposes, by comparing a relatively small amount of available truck-GPS data with a robust database of taxi-GPS data. The presented approach has potential to be used as an accurate travel time prediction method for several different business related applications or can be utilized by decision-makers for transportation planning purposes.

The analysis revealed that observed travel times from truck-GPS data can be successfully estimated by taxi-GPS data. Statistical calculations show that 60% of all AM period (6am-10am) commercial vehicle travel times can be estimated accurately using taxi data at the 95% confidence level. Among all time periods, the night period (7pm-6am) shows the greatest diversion between taxi and truck travel times, which most likely indicates that vehicular differences between taxis and trucks and their speeds are greater during closer to free-flow conditions. It is important to note that the differences in trip travel times are also affected by other factors such as road/lane restrictions for trucks, different route choices, and parking requirements. These facts are combined with the recurrent urban congestion in daytime and both taxis and commercial vehicles are exposed to significant delays.

Although the results of data analysis are sample specific, the steps presented for data analysis are applicable to similar datasets. Further improvements are possible using vehicle-specific weight factors to compensate for differences in vehicle types to enhance accuracy. Similarly, location specific weight factors can be included depending on the disturbances in traffic or to account for route restrictions. It should be also noted that travel time predictions for an entire network using a robust GPS data source can be a key component in developing or calibrating models and simulation-based studies.
REFERENCES


