ANALYSIS TOOL TO PROCESS PASSIVELY-COLLECTED GPS DATA FOR COMMERCIAL VEHICLE DEMAND MODELLING APPLICATIONS

Bryce W. Sharman, M.A.Sc.
Department of Civil Engineering
University of Toronto
35 St George Street, Toronto, Ontario, M5S 1A4, Canada
Tel: 416-978-5049; Fax: 416-978-5054; Email: bryce.sharman@utoronto.ca

Matthew J. Roorda, Ph.D. *
Department of Civil Engineering
University of Toronto
35 St George Street, Toronto, Ontario, M5S 1A4, Canada
Tel: 416-978-5976; Fax: 416-978-5054; Email: roordam@ecf.utoronto.ca

* Corresponding Author

INTRODUCTION

Many firms use GPS-enabled vehicle trackers to monitor their vehicle fleet. If such data can be accumulated for many firms it provides a very rich potential source of information to support modelling of the freight system for public sector decision-making. GPS data provide precise and continuous spatial and temporal information about a large number of vehicles for long periods of time. However, automated processing techniques are required to impute behavioural information about destination location and the frequency of repeated visits to destinations.

Current research efforts at the University of Toronto involve developing an agent-based model that simulates goods movement throughout an urban region (Roorda, et al., 2010). The *Logistics Decisions* component of this model will take as inputs a list of shipments for carriers in the region, will develop realistic shipment delivery schedules, assign the shipments to vehicles, and route the vehicles to the transportation network. Unlike most other travel demand models, this component will represent vehicle routing and scheduling over longer time periods to reflect daily, weekly or other rhythms of repetition that are apparent in delivery schedules.

High-quality data over periods of weeks or months are needed to estimate this travel demand model, as is the identification of repeated visits to customers and other destinations. Survey-based data collection for urban commercial vehicle travel is difficult, time consuming, and expensive (McCabe, et al., 2006). Due to the burden placed upon respondents, the duration of urban commercial vehicle surveys is usually limited to a single-day. Passive vehicle tracking using GPS has potential to increase the observation period of commercial vehicle surveys, while at the same time improving data quality and reducing respondent burden. With the decreasing costs and increasing availability of GPS units, commercial GPS surveys are becoming increasingly common for planning purposes.

Surveys that involve installing GPS units on the trucks of responding firms incur additional expenses, because of the cost of the technology and the effort required to install the device (generally requiring on-site visits). Given that a growing number of firms are already subscribing to continuous and longer-term GPS tracking services by third party providers, it is desirable to take advantage of these sources. Third-party data sources are already being used to measure truck travel times and speeds. For example, McCormack and Hallenbeck (2006) used GPS (among other forms of electronic data collection) to measure the effect of infrastructure improvements on truck travel times and speed on highways in Washington State. IBI Group (2008) has used GPS data to identify road congestion measures in the Quebec City-Windsor. Transport Canada uses GPS data to measure border wait times at Canada-US border crossing (Shallow, 2006). McCormack et al. (2010) used third-party commercially-available GPS data to support a state truck freight network performance monitoring program and to guide freight investment decisions by monitoring truck travel times and system reliability.

But GPS data from third-party sources are not yet being used as the sole source of data for estimating forecasting models of urban goods movement, except in limited explorations of trip generation, for example, for grocery stores (Bassok et al., 2010). This is largely because, while GPS vehicle-tracking information does provide accurate spatial and temporal information about truck movements, it does not provide behavioral information. Automated processing techniques are required to impute behavioral information about destination location and the frequency of repeated visits to destinations.

GPS data has been provided to University of Toronto researchers by a fleet management firm that places its custom GPS and vehicle engine monitoring devices on clients' vehicles. A proposed framework for the *Logistics Decisions* model component has been created that will use this passively-collected GPS data for model specification and estimation, including:

1. Hazard-based model for stop duration
2. Hazard-based model for the time interval between repeated visits to the same destination
3. Vehicle tour generation model showing the order in which different destinations are visited.
4. Model to assign shipments to different destinations to days within the study period.

This paper discusses the available GPS data and then highlights data processing techniques used to convert the stream of GPS data into a form suitable for travel demand modelling applications.

AVAILABLE DATA

Currently, GPS-recorded fleet management data has been provided by a third-party logistics firm for 40 firms between October 1st and October 31st, 2006. The data are intended for fleet-management applications and not intended as a travel survey or for demand-modelling applications. Hence, processing is required to convert this data into useable information for demand modelling purposes.

One issue is with the resolution of the GPS data. To save data storage and transmission costs, GPS data have only been recorded at 500 m intervals, with the distance between successive recorded GPS points increasing to every one or two miles as the vehicle reaches highway speeds.

Another issue is trip-end identification. Trip ends are automatically recorded by the vehicle-tracking unit when the vehicle remains stationary for a five-minute interval or if the engine is turned off. Analysis of the data showed that false-positive trip ends (recorded where no trip end should exist) are common, as approximately 11% of all recorded trip ends occurred on freeways. False-negative trip-ends were also found, where trip ends were not recorded even when the vehicle remained stationary for very long periods of time. Other criteria that can be used for trip-end detection are discussed in Greaves & Figliozzi (2008), Du & Aultman-Hall (2007) and Schuessler & Axhausen (2009) although many of these techniques cannot be used in the current dataset due to the resolution of the GPS points.

The study region is the Greater-Golden-Horseshoe (GGH) region in southern Ontario, Canada. This region is centered on Toronto (Canada's largest city) and Hamilton and forms the heart of Canada's industrial economy. Data were provided for vehicles within the study area. For vehicles entering the study region, vehicle points since the last stop before entering the study area were provided. Likewise data were provided for vehicles leaving the study area up until the first recorded stop.

DATA PROCESSING PROCEDURE

A custom program was created using C# to process the GPS data. This section contains a brief description highlighting the different aspects of this program, including: data cleaning, tracking vehicle GPS points, clustering GPS trip ends into destinations, and depot identification and tour creation.

Data Cleaning

Before analyzing the data, erroneous data are removed. Examples of poor data include infeasible latitude or longitude values. All stops are examined to test whether they are `false-positive stops`, meaning that it

is highly unlikely that vehicle visited a destination during this period. A stop is marked as a false-positive stop (and subsequently removed) if:

- It is located within a 40 m buffer of a freeway. This distance was chosen by visually inspecting the GPS points to ensure that the vast majority of vehicle tracks on the freeway remained within this buffer while minimizing the chance of a vehicle parked at a location near the freeway falling within the buffer.
- The stop duration was under 15 minutes and the stop was located within 20 m of a major arterial road. Since street parking is far more prevalent within downtown and urban regions than in suburbs and rural regions, this data-cleaning step was only performed outside of central Toronto.

Tracking Vehicle GPS Points

This section of the custom program follows the vehicle GPS points to identify stops and vehicle trips. This program also searches for missed (false-negative) stops, and removes short (within-yard) trips. GPS points are only recorded once the vehicle has travelled a certain distance. Missed stops are assumed if:

- The time interval between two subsequent GPS points exceeds 30 minutes.
- The average speed (time interval divided by the travelled distance between the two subsequent GPS points) is below 5 km/hr and at most one of the points is located on a freeway.

Short trips occur if the vehicle engine is turned on while not moving or only moving short distances, such as a repositioning trip within the truck yard. A short trip occurs if the straight-line distance between the origin and destination is less than 500 m or no GPS points are recorded between two subsequent stops (which would indicate the vehicle having travelled a minimum distance of 500 m).

Clustering GPS Trip Ends into Destinations

GPS recorded trip ends are independent events. The goal of this work is to group trip ends into *trip destinations*, which are distinguished from trip ends in that they reflect a business, household or other entity that attracts or generates the trip. Grouping (clustering) trip ends for an individual firm into repeatedly visited destinations allows shipment schedules to be inferred from GPS data, which is a crucial component of the proposed multi-day travel demand modelling framework.

Different clustering algorithms were tested in (Sharman & Roorda, 2010). That research showed that using Ward's clustering method, which is a hierarchical agglomerative clustering (HAC) method, produced the best results. This method is described in more detail in Kaufman, & Rousseeuw (1990) and Han, Lee, & Kamber (2009) and is summarized briefly below.

1. Create a distance (or dissimilarity) matrix showing the distance between any two points. In this application the distance is the Euclidean (straight-line) distance between two GPS points.
2. To start this process, every object is assigned to its own cluster.
3. In each step the closest two clusters are merged together. After merging, the distances between all clusters are recalculated. This step is repeated until all objects have been assigned to a single cluster

The hierarchical clustering tree can be cut at different heights (distances) or for a set number of clusters to produce different clustering results. The height at which the hierarchical tree is cut is called the *distance threshold* as it represents the largest distance within a cluster.

Finding the distance between two clusters is intuitive if each cluster only contains one object. When either of the clusters contains more than one object, however, there is no single and obvious distance measure. Ward's method is intended for interval-scaled measurements and Euclidean distances, which is the case in this research. The distance between two clusters is defined in equation 1.

$$d_w(R,Q) = \sqrt{\frac{2|R||Q|}{|R|+|Q|}\left\|\bar{x}(R) - \bar{x}(Q)\right\|^2}, \quad (1)$$

where $|R|$ represents the number of objects in cluster R, and *d(R,Q)* refers to the distance between clusters R and Q, *and* $\bar{x}(R)$ is the centroid of cluster *R*.

Selection of an appropriate distance threshold is difficult for this application. The first difficulty is that the distance measure between two clusters using Ward's method is increased by the number of points in each cluster, see equation 1. This causes Ward's method to produce tight and isolated clusters but makes specifying a distance threshold difficult for this application as there are reasonable physical limits to the size of a cluster. The other issue is the wide variation in land use, where properties vary on a scale from several meters up to very large properties extending hundreds of meters in a given direction. A large distance threshold that would join large numbers of objects in large properties is too large for other parts of the study domain. The use of a small clustering threshold, however, causes stops to a destination covering a large area to be split into multiple clusters.

The challenge of selecting appropriate clustering thresholds was overcome using a two-step process. In the first step, GPS trip ends were clustered using different distance thresholds up to a maximum threshold of 700 m. This step produces well-formed clusters that are not overly large. Many clusters may be formed, however, in large destinations where many trips ends are recorded over a large area. In the second step, GIS data were obtained identifying parcel boundaries for every property within the Greater Golden Horseshoe study region. The median object in every cluster is the object with the lowest sum-of-squared distances to other objects in the cluster. The cluster is assumed to belong to the parcel of its median object. Clusters belonging to the same parcel were merged into a single cluster.

Visual examination of the results proved extremely promising. The combined approach of clustering using tighter thresholds and then using land-use information to join the clusters on the same property provided tight separate clusters through the domain without artificially separating trip ends to destinations on large properties, which often occur in the truck yards of large trucking firms.

Property data were not available outside of the study region. In this case a large distance threshold was used to group the GPS trip ends as less precision is required for external stops.

Depot Identification and Tour Creation

This area of the program is currently under development. Criteria to determine whether a destination is a depot include the number of times that the destination was visited by all vehicles within the fleet and also if vehicles routinely remain at the destination for extended periods of time. Vehicle tours are created, by joining sequential trips observed in the data.

CONCLUSIONS

Many carriers already have GPS tracking devices placed on their vehicles to monitor their fleets, providing a broad source of information about the vehicle movements of many firms. This paper outlines the procedure used to process GPS data provided by a fleet management firm to make the data suitable for estimation of a demand model of commercial vehicle travel. A travel demand model has been proposed that uses this processed passively-collected GPS data from fleet management firms as its primary data source.

The main components of the data processing include data cleaning including the identification of false-positive stops. Then the GPS points are separated into different trips while cleaning within-yard trips and testing for missed stops. The trip ends are then grouped into their respective destinations (locations) using a two-step clustering approach. Finally work is ongoing into identifying depots and grouping the trips into trip chains (tours).

ACKNOWLEDGEMENTS

Financial support for this project came from the National Science and Engineering Research Council of Canada.

REFERENCES

Bassok, A., McCormack, E., Outwater, M., 2010. *Use of Truck GPS Data for Freight Forecasting: Toward Better Freight Transportation Data*. [Presentation] May 19, 2010. Puget Sound Regional Council (PSRC).

Du, J., & Aultman-Hall, L., 2007. Increasing the accuracy of trip rate information from passive multi-day GPS travel datasets: Automatic trip end identification issues. *Transportation Research Part A , 41*, pp.220-232.

Greaves, S., & Figliozzi, M.A., 2008. Commercial vehicle tour data collection using passive GPS technology: issues and potential applications. In: *CD Proceedings, 87th Annual Meeting of the TRB*, Washington DC, 13–17 January 2008.

Han, J., Lee, J.G., & Kamber, M., 2009. An overview of clustering methods in geographic data analysis. In: H. J. Miller, & J. Han, eds. 2009. *Geographic Data Mining and Knowledge Discovery.* 2nd ed. Boca Raton: CRC Press. Ch. 7.

IBI Group, 2008. *Continental Gateway Road Network Perfomance. Final Report Macro Analysis.*

Kaufman, L., & Rousseeuw, P.J., 1990. *Finding Groups in Data: an Introduction to Cluster Analysis.* NY: John Wiley & Sons.

McCabe, S., Kwan, H., & Roorda, M., 2006. Freight transportation: who is the decision maker? In: *53rd Annual North American Meetings of the Regional Science Association International*, Toronto, 16-18 Nov. 2006.

McCormack, E., & Hallenbeck, M. E., 2006. ITS devices used to collect truck data for performance benchmarks. *Transportation Research Record: Journal of the Transportation Research Board*, 1957, pp.43-50.

McCormack, E., Ma, X., Klocow, C., Currarei, A., Wright, D., 2010. Developing a GPS-Based Truck Freight Performance Measure Platform. WA-RD 748.1 (TNW 2010-02), [Internet] April 2010, Available at: http://www.wsdot.wa.gov/research/reports/fullreports/748.1.pdf. [Accessed 02-June 2010].

Roorda, M.J., Cavalcante, R., McCabe, S., & Kwan, H., 2010. A conceptual framework for agent-based modelling of logistics services. *Transportation Research Part E* , 46, pp.18-31.

Shallow, T., 2006. *Border Wait - Time Measurement Project. Presentation to the "Talking Freight Seminar Series" - August 16, 2006.* Washington, DC.: Federal Highway Administration.

Sharman, B., Roorda, M.J., 2010. Freight modelling with GPS data: a clustering approach for identifying trip destinations. In: *Proceedings of the TRANSLOG 2010 Conference* , Hamilton ON, 15–16 June 2010.

Schuessler, N. & Axhausen, K.W., 2009, Processing GPS raw data without additional information. In: *CD Proceedings, 88th Annual Meeting of the Transportation Research Board*, Washington DC, 11–15 January 2009.