A METHOD OF DATA INTEGRATION FOR GPS AND ROADSIDE INTERCEPT SURVEY DATA

Sirui Zhu, University of Toronto
Matthew Roorda, University of Toronto

Introduction

The Commercial Vehicle Survey (CVS) has been one of Ministry of Transportation Ontario's (MTO) main instruments for obtaining information about freight vehicle flows on the provincial transportation system. The CVS is a roadside survey that intercepts truck trips at data collection sites to gather information about inter-city truck flows as well as some urban truck flows. The survey is conducted every 5 years, and the most recent available data were collected in the period from 2010 to 2014, and is referred to as the 2012 CVS. During the survey, the intercepted truck drivers were interviewed to collect information about vehicle type, trip movement, and cargo contents. The 2012 CVS encompasses over 200 data collection sites and a total of 45,000 interviews (Ministry of Transportation Ontario, 2015). While the CVS is an extensive data collection effort, the sample collected through CVS is still a small portion of the entire truck population that flows through Ontario.

A new source of data is also becoming accessible to the MTO. Probe GPS tracking data from fleet management providers such as Shaw and ATRI have a greater geographic coverage of truck movement than intercept surveys, resulting in better representation of the truck population. It is estimated that GPS data cover as much as 50% of total vehicle kilometers travelled by trucks (Ministry of Transportation Ontario, 2015). GPS tracking data are generally comprised of detailed spot position (latitude and longitude) with time stamp for each individual truck at a fixed time interval. Connecting the position data in chronological order can yield a clear picture of the tour for each truck. However, the GPS tracking data are collected automatically by a device on the truck without any interaction from the driver, and the device lacks intelligence to collect data with more complexity such as cargo content. Relative to the cost of the CVS, the purchase of GPS tracking data is very inexpensive. GPS tracking data are currently used primarily for identification of congestion hotspots/bottlenecks (Zhao, McCormack, Dailey, & Scharnhorst, 2013). In summary, the CVS contains highly detailed information about truck flows at the points of intercept but lacks population coverage, while GPS tracking data provide a wider coverage of truck movements but lack useful supplementary information about the tour.

The purpose of this study is to apply data fusion methods (D'Orazio, Di Zio, & Scanu, 2006) to integrate these two data sets to produce a more useful combined source of information for modelling and policy analysis. The GPS data and the CVS data complement each other nicely in theory, but in practice the two sources of data have different levels of aggregation, sample methods, and statistical precision, all of which are common problems of data fusion (Polak, 2006). While challenges exist, the merits of data fusion are notable. First, it avoids the costly option of conducting an entirely new survey when the variables of interest exist in multiple previous surveys (Van Der Puttan, Kok, & Gupta, 2002). Second, it provides a means to analyze variables from different surveys within one platform (Bayart, Bonnel, & Morency, 2008). Third, it provides an approach to address the lack of comparability between data when a mix of survey methodologies are used (Bayart, Bonnel, & Morency, 2008).

This study divides the process of linking CVS attributes to GPS-based trips into two phases. The first phase is to identify all CVS records that share similar movements as the trips generated from GPS data. The second phase is to select the most appropriate CVS record for each GPS-based trip. The GPS-based truck trips will be enriched with information from the selected CVS record, including information such as vehicle configuration, body style, commodity type, commodity value, and weight attribute. This paper focuses on the first phase matching process with test results from sample data. Concepts for the second phase, which is still in progress, are presented at the end.

Literature Review

Data fusion has recently appeared in the transportation literature (Miller et al. 2012; Amey et al., 2009), but has been applied in other fields for a considerably longer time. Media research and consumer behavioural studies have used data fusion since 1980s (Van Der Puttan, Kok, & Gupta, 2002). To integrate two or more data sets in a data fusion procedure, the data sets must have the following characteristics (D'Orazio, Di Zio, & Scanu, 2006):

1. There must be a set of variables that are common across all data sets
2. Each data set must also have its own set of variables that are not observed by another data set
3. The units observed in each data set are different units from another data set

Statistical matching attempts to match records from different data sets based on their similarity in common characteristics rather than unique identification information (Rodgers & DeVol, 1984). A record in one data set can be matched to multiple records in a different data set, as long as the common variables are sufficiently similar. In statistical matching, there are two main approaches for data fusion: explicit modelling and implicit modelling. Explicit modelling (also called the classical approach) imputes the specific variables by creating a correlation model between the common variables and specific variables. Implicit modelling transfers values of specific variables from 'donor' records to 'receptor' records that have been matched on the basis of similar common variables (Bayart, Bonnel, & Morency, 2008).

A common imputation modelling technique is hot-deck imputation, which uses a procedure of selecting a donor record that shares the closest common variables with the receptor record and transfers specific variables from that donor record to the receptor record (Aluja-Banet, Daunis-i-Estadella, & Pellicer, 2007). The definition of 'closest' refers to the difference in values between the two records, and the method to calculate the difference can vary. Rudra et al. (2014) investigated many variations and modified forms of hot-deck imputation technique, such as random hot-deck, sequential hot-deck, and nearest neighbour donor.

GPS Data Preprocessing

The truck GPS data used in this study was provided by Ministry of Transportation Ontario. The data describe the movement of freight vehicles over the course of one month through a series of latitude and longitude coordinates with time stamp and unique ID for each truck. The GPS data include points both when the vehicle is moving and when it is stopped. Thus, preprocessing of GPS data is necessary to filter out the GPS points of interest. The processing procedure distinguishes trip ends at a facility location from traffic stops on the roadway, stops at rest area on highway, and stops at border custom check. This study uses similar standards of trip end identification applied by Sharman (2014), who used similar data from RouteTracker^{TM} GPS units to model activity behaviour and inter-arrival duration of urban commercial transportation. Any non-commercial related stops at highway rest area and border custom check area are

removed through analysis in a Geographic Information System. The resulting trip ends form the origins and destinations of GPS trips. From these GPS trips, any trip movement whose origin and destination is within the United States only is also discarded, because the CVS only contains trip information relevant to Canada. The resulting trip origins and destinations were visually inspected on GoogleMap™ to ensure accuracy of preprocessing.

A preliminary test for the study used GPS data for 36 vehicles to establish truck movements over the course of one month. In this one month, 36 vehicles had 2002 stop points, which generated 1315 trips. 491 of these trips had either an origin or destination in Canada.

Data Fusion Procedure

This study uses a two-stage procedure to fuse the GPS data and the CVS data. The first stage is to identify candidate matches between GPS trips and CVS records. The second stage is under development and will select a CVS record from the pool of candidate matches for each GPS trip.

In order to be considered a candidate match, the origin and destination of trip from a CVS record must be the same as the origin and destination of GPS trip. The spatial criterion for comparing origins and destinations between CVS and GPS can be relaxed or tightened. The spatial scales considered in this study, for Canada, are census division (CD), census subdivision (CSD), census dissemination area (CDA), and exact location within a 200m threshold. For the US, two spatial scales are considered: County and exact location within a 200m threshold. For example, when the spatial scale of CSD is used, the origin of a CVS record is considered a match for the origin of GPS trip if both are located within the same CSD. When both origin and destination of a CVS record are matched to a GPS trip origin destination at a particular spatial scale, then this CVS record becomes a candidate match for the GPS trip at that spatial scale. Note the term ‘candidate match’ is used because the GPS trip can have multiple matches from the CVS data.

It becomes more likely for CVS records to qualify as candidate matches when a more aggregate spatial scale is used. This increases the proportion of GPS trips that have candidate, but at same time generates candidates for each GPS trip that may be of lower quality (matching locations at the 200m threshold is clearly preferable over matching at the census division level). The matching criteria between CVS and GPS trips should be relaxed to increase the number of GPS trips that have matches, but the relaxing of matching criteria should be implemented sequentially to control the number and quality of candidates.

A further complication is that several stops are reported in a CVS interview. In addition to the trip origin and trip destination, the CVS records also have information on the previous stop before the data collection site (DCS) and next stop after the DCS. The previous stop and next stop are intermediate stops of a multi-stop tour, stopping at home, reporting to head office, or any other planned stops that are not traffic related. In most cases the previous stop is same as trip origin and the next stop is same as trip destination, but for some records the stops are different. It is possible that origin-destination of GPS trip match some other stop combinations of CVS records. There are five other combinations other than the standard trip origin to trip destination combination (Figure 1).

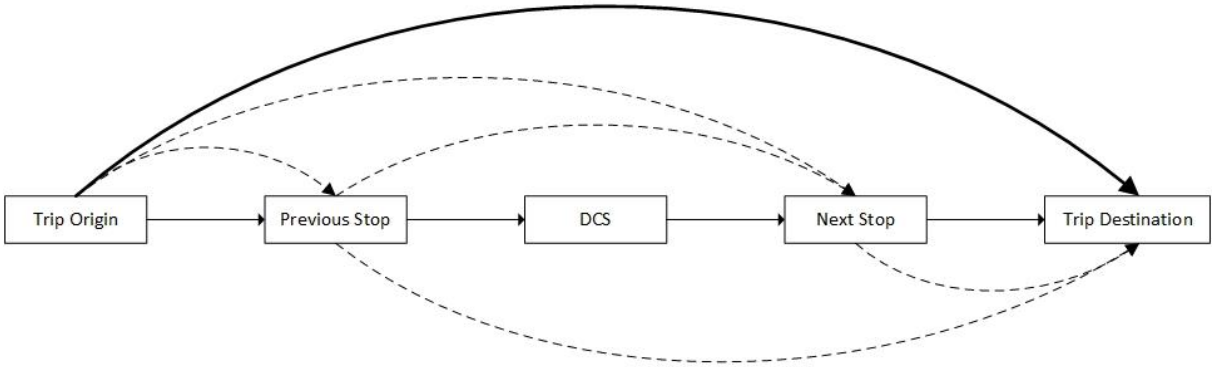

Figure 1 – Possible combination of stops from CVS records

The first stage procedure begins with the most detailed spatial scale: the exact location with 200m threshold. It searches the various combinations of stops, beginning with the trip origin and destination, and then considering other combinations shown with dotted lines in Figure 1. If a match has not been found, a more aggregate spatial scale is selected, and various combinations of stops are re-tested. The procedure is conducted separately for the GPS trip origin and trip destination; so the GPS trip origin could be matched to a CVS record at the CDA spatial scale, but the GPS trip destination is matched to the same CVS record at the CD spatial scale.

The outcome of the first stage is that each GPS trip is matched with a pool of candidate CVS records. The procedure captures all possible matches from CVS records for each GPS trip. A summary of the first stage matching outcomes using various cases of spatial scale and stop combination criteria is shown in Table 1. With greater spatial aggregation, it is more likely for each GPS trip to have at least one candidate match from CVS records, and a greater number of matched records. At the most detailed level, where trip origin and destination of the CVS trip are required to match within 200m of the GPS origin and destination (Case 1), only 6% of the GPS trips are matched with a CVS record, and most only find a single matching record. At the other extreme, when the matching is done at census division spatial scale for both trip origin and trip destination (Case 5), 93% of GPS trips found at least one candidate match from the CVS records.

Compared to Case 1, Case 2 relaxes the stop combination criterion of stop combination, such that other combinations of CVS stops (as shown in Figure 1) can be matched. This relaxation allows for an additional 1% of GPS trips to be matched to a CVS record.

Table 1 – Results of First Stage Matching Procedure

Case	Stop Combination Criterion	Spatial Scale Criterion	% of GPS trips with match	Avg no of candidates per matched GPS Trip
1	CVS Origin and Destination Only	CVS Origin and Destination match GPS within 200 m threshold	6%	1.107143
2	All Combinations	CVS Origin and Destination match GPS within 200 m threshold	7%	1.096345
3	CVS Origin and Destination Only	CVS Origin and Destination match GPS at Census Dissemination Area	54%	5.950570
4	CVS Origin and Destination Only	CVS Origin and Destination match GPS at Census Subdivision	79%	23.843187
5	CVS Origin and Destination Only	CVS Origin and Destination match GPS at Census Division	93%	69.665938

The second-stage procedure is under development. It will select the most appropriate CVS record from the pool of candidate matches. The selected CVS record would be considered 'fused' to the GPS trip, and attribute information of that CVS record would be assigned to the GPS trip.

The selection of the CVS record to fuse to the GPS trip will account for the quality of the match using a penalty system. A penalty is assigned for each candidate match CVS record depending on the spatial scale and stop combination used. The penalty recognizes that:

a) Relaxing the spatial scale criterion incurs a higher penalty (i.e. we prefer to match records at a fine spatial scale)
b) Different CVS stop combinations may incur higher penalties (e.g. matching a CVS record on the basis of its trip origin and destination is preferable to matching on the basis of its trip origin and next stop)

The selection of the CVS record will account for the weight assigned to the CVS trip. CVS trips that have been assigned a higher weight (via the MTO vehicle sampling approach) should have a greater likelihood of being selected.

The selection of the CVS record will also account for correlation among multiple consecutive GPS trips on a tour. Clearly, consecutive trips on a tour are made by the same vehicle configuration, and would normally be carrying a similar commodity class. Such criteria will further inform consistent CVS record selection for trips within a tour.

Finally, there will be many GPS records for which there is a pool of potential candidate matches. We plan to select an individual record randomly from within that pool based on probabilities that account for penalties and constraints described above (to increase the probability of selecting a higher quality match).

Conclusion and Next Steps

A two-stage procedure is developed to fuse data from the MTO Commercial Vehicle Survey with GPS vehicle tracking data. The first stage identifies candidate matches between GPS and CVS records. The second stage selects a CVS record from the pool of candidate matches for each GPS trip.

Our preliminary results for the first stage procedure indicate that there are matching truck movements between CVS data and GPS data, sometimes at a very fine spatial scale. However, to increase the number of GPS trips that can be matched to a CVS data record, the matching criteria should be relaxed. To reflect the quality of the match, a penalty system prioritizes matching candidates at disaggregate spatial scales.

The second phase will use penalty scores produced in the first phase, sampling weights assigned to the CVS data, GPS tour characteristics and a random selection procedure for the GPS - CVS data fusion. Once each GPS trip is fused with a CVS record, the study will proceed to evaluate the quality of the fused data through external and internal validation exercises.

References

Aluja-Banet, T., Daunis-i-Estadella, J., & Pellicer, D. (2007). GRAFT, a complete system for data fusion. *Computational Statistics & Data Analysis*, 635-649.

Amey, A., Liu, L., Pereira, F., Zegras, C., Veloso, M., Bento, C., & Biderman, A. (2009). *State of the Practice Overview of Transportation Data Fusion: Technical and Institutional Considerations.* MIT-Portugal Program.

Bayart, C., Bonnel, P., & Morency, C. (2008). Survey Mode Integration and Data Fusion: Methods and Challenges. *8th International Conference on Survey Methods in Transport.* Annecy, France.

D'Ambrosio, A., Aria, M., & Siciliano, R. (2007). Robust Tree-based Incremental Imputation method for Data Fusion. *7th International Symposium on Intelligent Data Analysis.* Ljubljana, Slovenia.

D'Orazio, M., Di Zio, M., & Scanu, M. (2006). *Statistical Matching: Theory and Practice.* New York: John Wiley & Sons.

Gilula, Z., & Rossi, P. (2004). A Direct Approach to Data Fusion.

Miller, E., Nurul Habib, K., Lee-Gosselin, M., Morency, C., Roorda, M., & Shalaby, A. (2012). *Changing Practices in Data Collection on the Movement of People: Final Report.* Sainte-Petronille: Lee-Gosselin Assciates Ltd.

Ministry of Transportation Ontario. (2015). *Commercial Vehicle Survey Expansion Factors.*

Ministry of Transportation Ontario. (2015). *MTO - UofT GPS Probe Data Research Initiative.*

Polak, J. (2006). *OPUS Final Report.* London: Imperial College London.

Rodgers, W., & DeVol, E. (1984). An Evaluation of Statistical Matching. *Journal of Business & Economic Statistics, 2*(1), 91-102.

Rudra, M. (2014). *Application of Imputation Methods in the Analysis of Freight Trip Generation in the Greater Toronto and Hamilton Area.* Toronto: University of Toronto.

Scheuren, F., & Winkler, W. E. (1993). Regression analysis of data files that are computer matched. *Survey Methodology*(19), 39-58.

Sharman, B. W. (2014). *Behavioural Modelling of Urban Freight Transportation: Activity and Inter-Arrival Duration Models Estimated Using GPS Data.* Toronto: University of Toronto.

Van Der Puttan, P., Kok, J., & Gupta, A. (2002). *Data Fusion Through Statistical Matching.* Cambridge: Center for eBusiness@MIT.

Venigalla, M. (2004). Household Travel Survey Data Fusion Issues. *National Household Travel Survey Conference.* USA.

Winkler, W. E. (2014). *Matching and Record Linkage.* U.S. Bureau of the Census. ResearchGate.

Zhao, W., McCormack, E., Dailey, D., & Scharnhorst, E. (2013). Using Truck Probe GPS Data to Identify and Rank Roadway Bottlenecks. *Journal of Transportation Engineering, 139*.