EXPANSION OF A GPS TRUCK TRIP SAMPLE TO REMOVE BIAS AND OBTAIN REPRESENTATIVE FLOWS FOR ONTARIO

Kevin Gingerich, Cross-Border Institute, University of Windsor
Hanna Maoh, Cross-Border Institute, University of Windsor
William Anderson, Cross-Border Institute, University of Windsor

Introduction

A prominent method of data collection for passenger and freight transportation is the application of Global Positioning System (GPS) devices to capture vehicle movements. This trend is the result of a successful integration of GPS technology as a commonplace occurrence for navigation. Moreover, commercial firms employing fleet vehicles for goods movement widely adopted GPS to remotely observe their vehicles and deploy their resources accordingly. The data generated from these GPS devices result in data pings - individual spatiotemporal points that identify where a vehicle was located over space at a particular instance of time. These pings also typically denote a particular vehicle and company using some form of identification, though this is often anonymized to protect the identity of the drivers and firms. Using the identification information, GPS pings for a particular truck can be combined together to observe the movements of the vehicle over time and convert this information into vehicle trips.

The large output of data generated from GPS devices, and the increasing availability of such information from vendors such as the American Transportation Research Institute (ATRI), are positive features that position the GPS data source as a viable alternative to traditional data collection models. Moreover, the data generated from GPS pings avoid potential recall errors that may occur from surveys requiring a respondent to reconstruct past activities (Stopher and Greaves, 2007). However, this data source also carries some potentially negative drawbacks. The data does not typically provide explicit information on the nature of activities carried out by trucks. This is common for GPS data derived from commercial fleets with no expectation of their data adapted later for transportation models. By comparison, there are examples of data surveys that are based on travel diaries to supplement the GPS data with activity information (Du and Aultman-Hall, 2007). Such surveys are typically designed from the ground up for modelling purposes and require a higher level of involvement from the survey respondent. In turn, this can lead to additional compensation to the respondent, thereby increasing the cost of the survey and typically limiting it to a lower number of participants.

Post-processing methods on passive GPS datasets obtained from fleet tracking companies can overcome the high costs associated with devising specialized surveys. For example, Gingerich et al. (2016a) employed entropy as a method of differentiating the purpose of stops as primary (to transfer goods) or secondary (for truck driver/vehicle needs). For passenger trips, Bohte and Maat (2009) used vehicle speeds and spatial proximity to geographic features to determine the trip purpose and mode of transportation. These emerging efforts are promising as they can be used to analyze the patterns inherent in the data and utilize other spatial information to infer the activity patterns. However, the application of different methods on such data has to consider the potential bias resulting from a non-representative sample of vehicles in the GPS dataset. Failure to address this concern will lead to erroneous conclusions about the movement patterns. Bricka et al. (2009) discovered differences among the demographic characteristics of responders for traditional surveys and GPS based surveys used to observe household travel patterns. For commercial vehicles, a GPS dataset often comes from a single GPS service provider that is employed by one or more commercial companies to supply the devices and software necessary to track a fleet of vehicles. The set of companies that use a single service provider are more likely to be interested in similar services since each service provider may provide differentiated products. Moreover, the characteristics of the firms that require a GPS service provider may

differ from other firms that do not require such extensive tracking services. For example, trucks with local routes or consistent schedules will not necessarily require the resources of a GPS service provider.

This paper addresses the non-representation issue in GPS truck data by offering a procedure for expanding a sample of truck trips traveling between census divisions in Ontario. The contribution of our work is three-fold. First, the sample of data can be extended to match aggregate totals of truck trips in Ontario. Secondly, biases observed in the data can be accounted for during the expansion. Finally, the resulting expansion factors can provide a measure of freight trip generation that is easily transferable across Canada.

The rest of this paper begins by describing the GPS dataset used as a case study and the biases encountered in the data. Next, a novel approach is provided to reduce the issue of bias in the dataset and expand the data sample to match aggregate totals. The methods and validation results of this case study are provided before closing the paper.

Primary Data and Biases

A GPS dataset loaned by Transport Canada forms the basis of analysis in this study. A subset of the data pertaining to the month of January, 2013 was utilized. The January data pertains to approximately 730 Canadian owned trucking firms and 40,000 individual trucks. Processing the GPS dataset produced 250,000 trips representing trucks that travel within Canada and across the U.S. These trips were derived from inter-zonal truck movements between census divisions (Canada) and metropolitan statistical areas (MSA)/ counties (U.S). A detailed description of the processing used to derive these trips can be found in Gingerich et al. (2016b). This includes the processing of vehicle stop events, classification of stops as primary or secondary, generating trips between primary stop events, and development of a time based constraint to determine an allowable travel time for a reasonable trip. In addition, the industry of the trip is estimated based on the nearest firm (within 200 meters) to the stop location. The list of approximately 507,660 Canadian firms included their location and industry. The latter data was purchased from InfoCanada for the year 2014. Since each trip is bounded by a primary stop at each end of the trip, an origin industry and destination industry are both estimated.

An analysis of the GPS data led to the discovery of several biases inherent in the sample of trips. For example, a comparison of the GPS data with the 2006 MTO commercial vehicle survey (CVS) data was performed. In Ontario, both datasets utilized the same zones at a census division level. Outside of Ontario, the MTO dataset utilizes larger zones by aggregating to the province/state level. A distribution of the trips from each dataset was created based on the distances between zones as shown in Figure 1 (bins of 400 km were used). The data was standardized for comparison purposes by calculating the total proportion of trips for a given distance range. In both cases, the frequency is highest for short range trips and reduces with an increasing distance. In fact, the MTO CVS dataset provides a fairly smooth curve that would fit well with negative exponential or power curves often associated with gravity models of trip distribution. By comparison, the distribution for the GPS dataset shows a lower proportion of short distance trips, while longer trips exhibit a higher overall proportion beginning with trips traversing more than 800 km in length. The higher proportion of longer distance trips matches our expectations discussed in the introduction since trucks/firms with short range trips are less likely to rely on a GPS service provider.

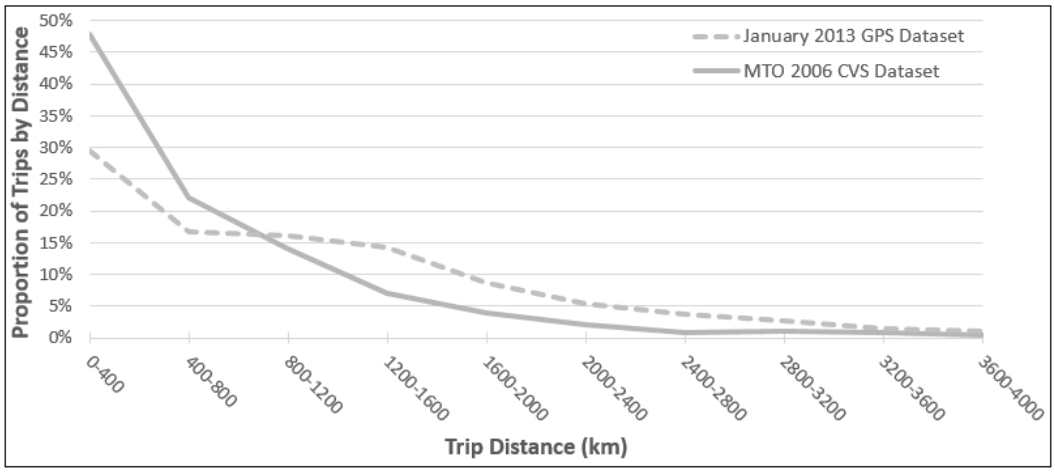

Figure 1: Distribution of trips grouped by distance

Industry biases is another issue in the representation of the GPS dataset. In Ontario, trips derived from the GPS data only utilize 5 different mining firms. By contrast, the firm dataset for Ontario purchased from InfoCanada estimates a total of 988 mining firms that exist in Ontario. Our sample therefore covers only 0.5% of these firms. By comparison, all categories of industry are represented by 9,097 firms in the GPS dataset and 507,660 firms in the InfoCanada firm database for a sample proportion of 1.8%. The proportional representation of firms by industry for the sample of GPS derived trips is provided in Figure 2. Manufacturing and Transportation exhibit higher proportions of representation, while primary industries ('mining' and 'agriculture, forestry and fishing') and services contain a lower proportional representation. For the service industry, the shown result may be intuitive since not all service firms will require commercial trips from commercial trucks. However, a larger representation is expected for primary industries where goods distribution is more common.

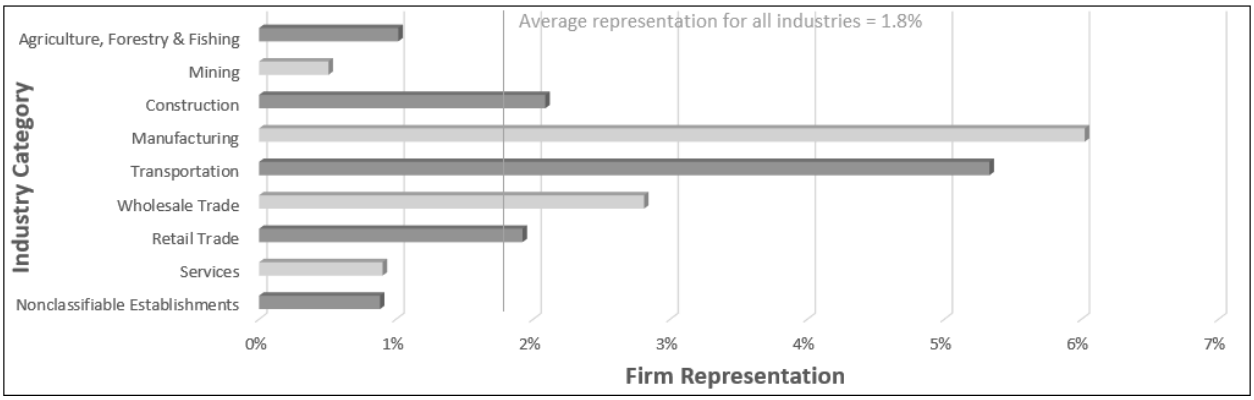

Figure 2: GPS Sample Representation of Firms by Industry

The lower representation of primary industries may be the result of the method used to estimate the industry of a trip end. A point to point relationship was established between the stop event location and the nearest firm within 200 meters. The point of a firm is often located at the road entrance to the property since the address of the business is generally used for geocoding. If a property is extremely large, as in the case of many primary industries where large land space is required, the actual point for the firm may be located outside the search radius and remain undetected. Utilizing lot boundary information is a potential method of mitigating this issue, but can be difficult to obtain. This is particularly true when observing our original dataset across Canada and the U.S, where individual lots with business information would need to be obtained from each municipality independently.

Methods and Results

Based on the analysis described above, we identified two major types of bias with the GPS dataset: (1) a spatial bias where our dataset over-represented longer distance trips and (2) an industry bias where primary industries are particularly under-represented. This section describes the methods that were devised to reduce this bias while also expanding the sample data to match aggregate totals as shown in Figure 3. The numbers in Figure 3 represent the order of each step, and are used as a reference for the remainder of this section of the paper.

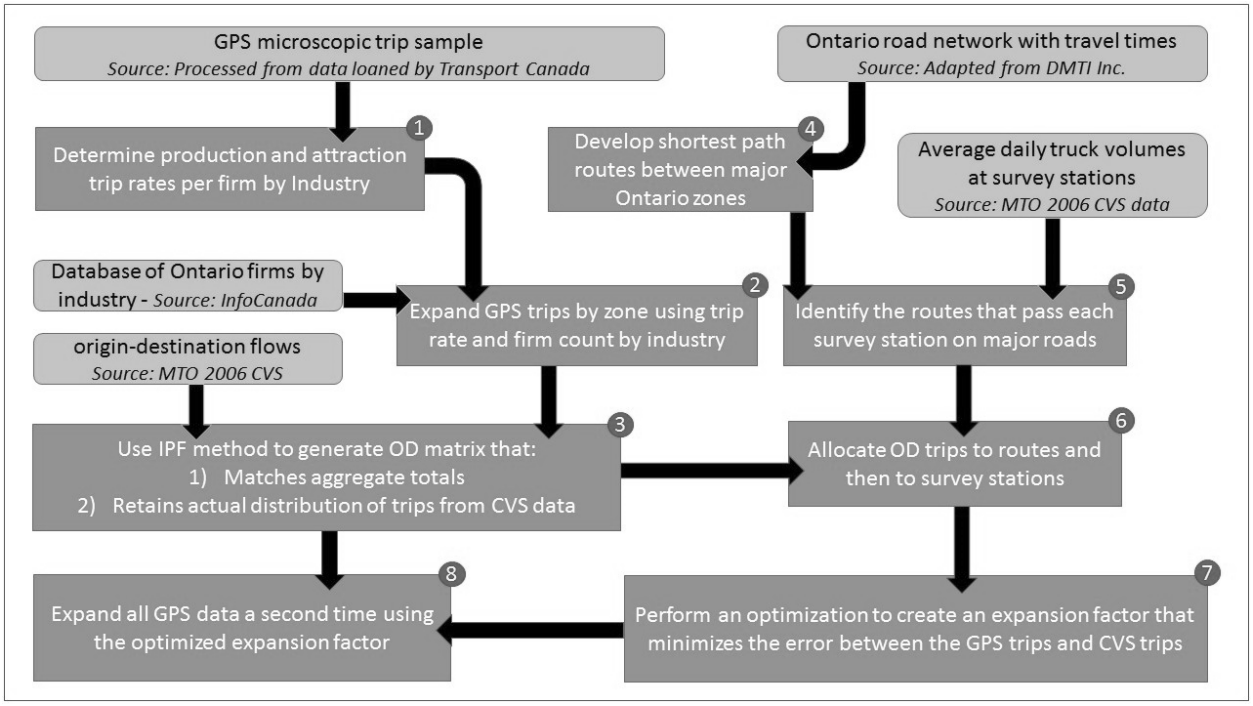

Figure 3: Flow chart outlining methods of analysis

Trip Rates and Initial Expansion

Step 1 - To begin, we have estimated average trip rates per firm over one month calculated as:

$$R = \frac{T}{F}$$

where *R* is the trip rate, *T* is the number of trips, and *F* is the number of firms. Average trip rates for Ontario, the rest of Canada, and the U.S, are provided in Table 1 for trip productions and attractions. The results show that the trip rates for Ontario are the highest. The rest of Canada exhibits a slightly lower rate compared to Ontario while the U.S exhibits a substantially reduced trip rate. The large drop in trip rate for U.S firms is likely caused by the nature of the GPS data source tracking only Canadian owned carriers. As a result, U.S firms are visited less frequently. The small difference between Ontario and the rest of Canada can be attributed towards the larger density of activities in Ontario compared to some of the other Canadian provinces.

Table 1: Average Trip Rates by Jurisdiction

Jurisdiction	Production			Attraction		
	Trips	Firms	Trip Rate	Trips	Firms	Trip Rate
Ontario	56,423	9,097	6.20	54,965	9,056	6.07
Rest of Canada	83,373	14,153	5.89	83,116	14,373	5.78
U.S	58,507	25,161	2.33	59,716	25,479	2.34

While the case study focuses specifically on Ontario, the Ontario trip rates exhibited issues caused by small sample sizes since the primary industries are under-represented (as shown previously in Figure 2). For all other industries, the trip rates for Ontario and the rest of Canada were similar. Since we are interested in providing a better representation of industry categories across the dataset, we utilized the Canadian trip rates by industry. The results of the Canadian trip rates are provided in Table 2 below.

Table 2: Average Trip Rates for Canadian Firms by Industry

Industry	Production			Attraction		
	Trips	Firms	Trip	Trips	Firms	Trip Rate
Agriculture, Forestry & Fishing	1,498	285	5.26	1,466	282	5.20
Mining	778	133	5.85	816	146	5.59
Construction	9,697	1,925	5.04	9,827	1,883	5.22
Manufacturing	23,900	3,821	6.25	23,556	3,784	6.23
Transportation	38,173	3,691	10.34	37,770	3,815	9.90
Wholesale Trade	12,416	1,931	6.43	12,461	1,957	6.37
Retail Trade	22,865	4,906	4.66	22,333	4,979	4.49
Services	22,291	4,956	4.50	22,116	5,036	4.39
Total	**139,796**	**23,250**	**6.01**	**138,081**	**23,429**	**5.89**

The total trip rate was also examined at the zonal (Ontario census division) level to determine if the trip rates are consistent over space. Figure 4 presents the relationship between the number of trips in the GPS sample and the subsequent trip production rate for each Ontario census division. It should be noted that five outliers (out of 49 zones) were removed from the plot, including three points with low total trips but very high trip rates (above 12) and two points with very high total trips but reasonable trip rates in line with the curve in Figure 4. The trend line in Figure 4 suggests that as the number of trips encountered for a given zone increases, the trip rate generally increases as well. However, the relationship itself is non-linear – as the number of trips increases, the trip rate increases at a slower pace. The trend suggests a general convergence of the trip rates approaching 7 trips per firm. It also suggests that our sample derived trip rate may under-estimate the actual trip rate of firms. The under-estimation issue is handled through an optimization approach as will be described later on in this paper.

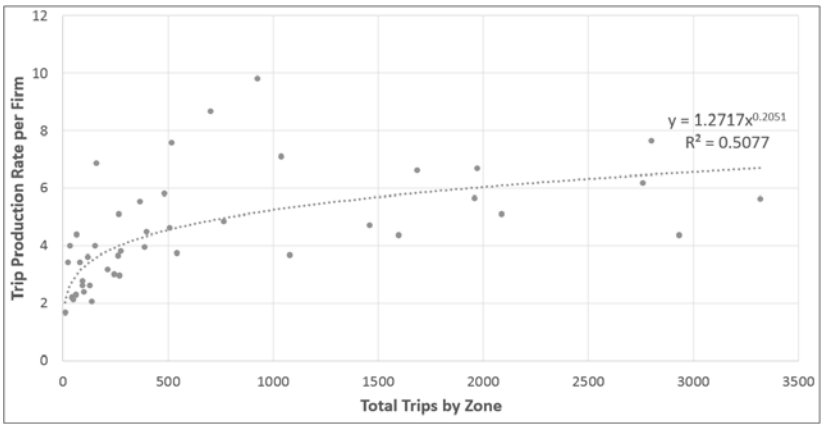

Figure 4: Total trip count and trip rate relationship for Ontario census divisions

Steps 2 and 3 - The trip rates per firm by industry given in Table 2 were multiplied by firm counts at the zonal level. This created an expanded aggregate trip total that adjusts the industry proportions based on the frequency of firms in each zone. The result of this expansion is a set of production and attraction totals per zone by industry. Since the distribution of trips between origin and destination was previously found to be biased towards longer distance trips in the GPS dataset, this pattern was not utilized to disaggregate the production and attraction totals. Instead, the pattern of distribution from the MTO 2006 CVS was used to

create the origin-destination matrix. The iterative proportional fitting method (IPF) was applied to update the origin-destination matrix. This method matches the expanded production and attraction totals while preserving the underlying spatial interaction pattern derived from the CVS data.

Trip Optimization and Second Expansion

Step 4 - To determine the suitability of the data obtained from the initial expansion, we compare the resulting traffic flows of trucks with point survey data along major highways (based on MTO 2006 CVS survey stations). Moreover, we can use these actual totals to appropriately adjust the GPS derived aggregate totals a second time. We estimated the traffic flows emerging from the origin-destination results based on an all-or-nothing traffic assignment (i.e. using shortest travel time) between each of the 49 zones as shown in Figure 5. The free flow travel time was used for this purpose since the primary truck routes between these zones are typically large capacity highways that will also be utilized under congested conditions.

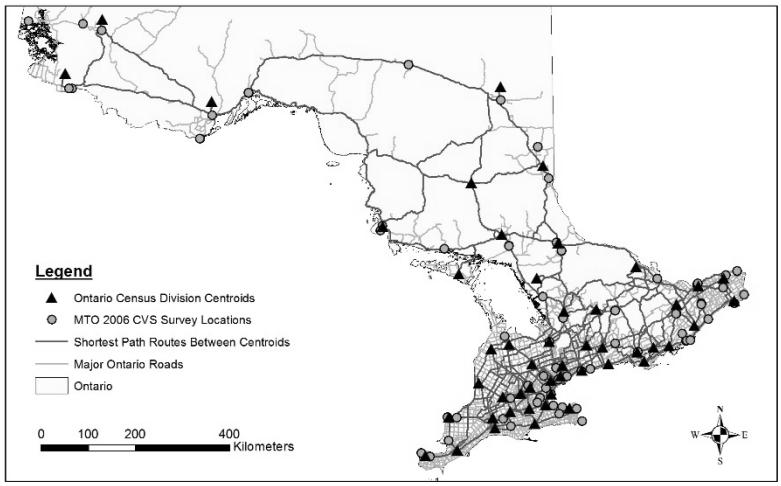

Figure 5: Shortest path routes between Ontario census divisions and MTO 2006 CVS survey stations

Steps 5 and 6 - A relationship was developed between the shortest path routes for the 49 census divisions and the MTO survey points located across Ontario. This was done by determining the routes that pass along each survey point. Using this relationship, the origin-destination trips were assigned to the appropriate routes (i.e. an AON assignment is performed). Next, the AON traffic volume was further assigned to each survey station for comparison. The results of this comparison indicate that 77% of the total trips at the CVS survey stations are accounted for by the current expanded totals of the GPS sample trips.

Steps 7 and 8 - The current totals from the GPS sample can be further expanded to match the trips observed by the CVS survey stations. To accomplish this task, a non-linear optimization problem is formulated, where the objective function minimizes the total error between the CVS survey station totals and the traffic flows derived from the expanded GPS trip totals. As such, a single weighting multiplier value is introduced to adjust all GPS trip totals simultaneously. The optimization is designed as follows:

Minimize: $\varepsilon = \sum_s^n |t_{s,CVS} - w t_{s,GPS}|$

Subject to: $w \geq 0$

where ε is the total error to be minimized and w is the variable multiplier adjusted in the algorithm. $t_{s,CVS}$ and $t_{s,GPS}$ are the trip totals of survey station s from the CVS data and the GPS data, respectively, for all s=1, 2,…, n (n = 45) survey stations located on at least one shortest path route. The optimization resulted in a final multiplier value of 1.27. This multiplier expands the origin-destination data derived from GPS trips a second time to reach a final total that corresponds to actual traffic totals as closely as possible. A scatterplot showing the final CVS totals and expanded GPS totals is provided in Figure 6. The graph

indicates a very strong one-to-one relationship between the two trip sets with a linear trend line slope of 1.01. Furthermore, the correlation between the two sets of data is 93.9%. A map of the errors suggested by Figure 6 are plotted in Figure 7. This map shows that the Toronto and Hamilton areas exhibit a higher actual total measured from the CVS data, while areas primarily north east of Toronto experience higher GPS totals. This can be expected due to the potential for intra-zonal trucks that are not accounted for by the GPS trips.

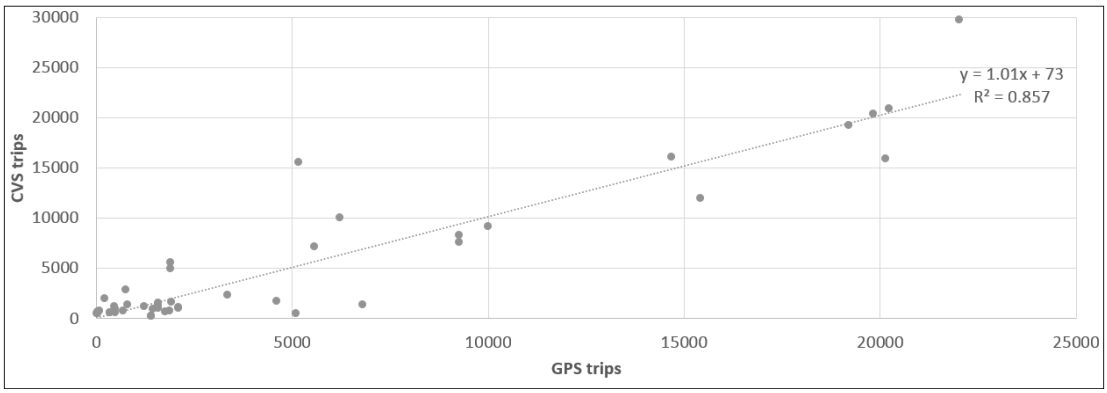

Figure 6: Relationship between the optimized GPS trip totals and observed CVS trips per survey station

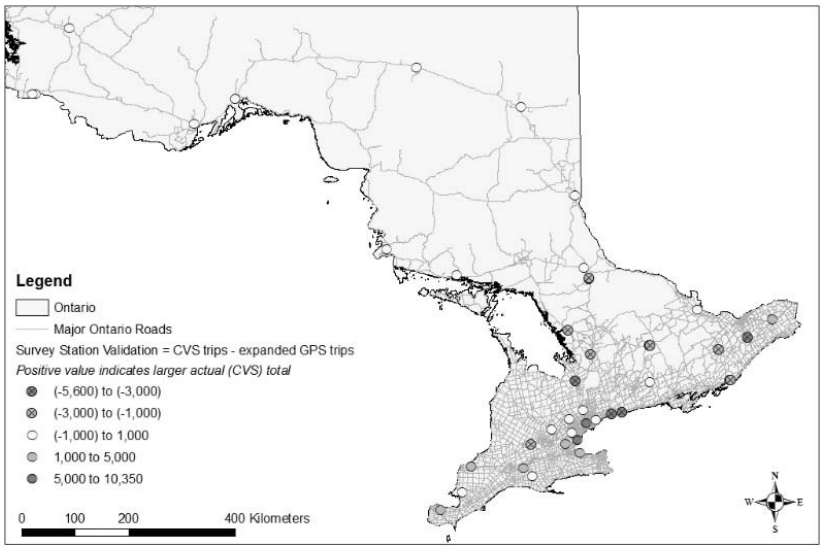

Figure 7: Validation results comparing trip totals at the Ontario survey stations

Conclusions and Future Work

This paper identified two types of bias, industry and distance, found in a sample of GPS derived truck trips. A method was established to remove the industry bias using trip rates and expanding by the population of firms in a given zone. In addition, distance bias was accounted for by utilizing the IPF method to match total estimated zonal production and attraction (from the first expansion) while maintaining the origin-destination patterns obtained from the 2006 CVS survey created by MTO. A second expansion was then applied by optimizing the expanded GPS totals with the truck totals from survey station points located along major Ontario routes.

While we used 2006 CVS data for the second expansion, 2013 data has been prepared by MTO. Based on a simple comparison of the two datasets, we expect a truck trip increase of 40% between Ontario census

divisions (102,175 trips in 2013 compared to 72,870 trips in 2006). In such a case, the 2006 multiplier value of 1.27 may be increased by 40% to 1.74. However, the final value may be slightly different if the 2013 trip counts did not increase at each location proportionally. We contend that the utilization of a single multiplier value for the second expansion has a clear advantage since we can merge the original expansion factors (trip rates by industry) with the optimized factor from the second expansion. For example, the manufacturing production trip rate of 6.25 per firm (from Table 1) and the second expansion factor of 1.27 would become 7.94 (6.25 × 1.27). The simplicity of a single factor for each industry type ensures that these trip generation rates are easily applicable in the Canadian context.

The total trip productions and attractions generated from the analysis provided a better representation of truck trips in Ontario compared to the original sample while closely matching the aggregate totals observed on the road network. However, the microscopic behaviour of individual trips is lost at an aggregate level. To retain the travel behaviour of vehicles, we plan to utilize the original sample to synthesize a full population of trips by using methods such as combinatorial optimization (Ryan et al., 2009). In such a case, the synthesis algorithm can be used to ensure that the aggregate zonal totals by industry type are maintained. Such a method has been applied before for expanded trip rates. For example, Goulias et al. (2014) used population synthesis to expand a household survey in California. After the trips are synthesized, our data can then be used in microscopic transportation models (such as truck tours) without the biases inherent in the original GPS sample.

Acknowledgements

The authors would like to thank Louis-Paul Tardif and Andrew Carter from Transport Canada for loaning the GPS dataset to us. In addition, we would like to thank Zachary Stephen for his help processing the data. Finally, we are also grateful for the financial support provided by FedDev Ontario and NSERC.

References

Bohte, W., and Maat, K., 2009. Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands, *Transportation Research Part C*, vol. 17, pp. 285-297.

Du, J., and 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*, vol. 41, pp. 220-232.

Gingerich, K., Maoh, H., and Anderson, W., 2016a. Classifying the purpose of stopped truck events: An application of entropy to GPS data, *Transportation Research Part C*, vol. 64, pp. 17-27.

Gingerich, K., Maoh, H., and Anderson, W., 2016b. Characterization of International Origin-Destination Truck Movements across Major Ontario-Michigan Border Crossings, *Transportation Research Record*, In Press.

Goulias, K., Ravulaparthy, S., Konduri, K., and Pendyala, R., 2014. Using synthetic population generation to replace sample and expansion weights in household surveys for small area estimation of population parameters, *Compendium of the Transportation Research Board 93rd annual conference*.

Prozzi, J., Wong, C., and Harrison, R., 2004. Texas truck data collection guidebook, published by the Center for Transportation Research, The University of Texas, Austin, pg 21-24.

Ryan, J., Maoh, H., and Kanaroglou, P., 2009. Population synthesis: Comparing the major techniques using a small, complete population of firms. *Geographical Analysis*, vol. 41, pp. 181–203.

Stopher, P., and Greaves, S., 2007. Household travel surveys: Where are we going?, *Transportation Research Part A*, vol. 41, pp. 367-381.

[1] Type of Paper: Regular