# THE FLUIDITY OF THE CANADIAN TRANSPORTATION SYSTEM: A COMMERCIAL TRUCKING PERSPECTIVE

Alexander Gregory[1], Transport Canada
Kristina Kwiatkowski[1], Transport Canada

## Introduction

As part of its effort to measure the efficiency of the transportation system and Canada's Gateways and Corridors, the Economic Analysis directorate at Transport Canada has been developing reliability indicators to measure and benchmark the Canadian transportation system. These indicators are intended to quantify the reliability of the supply chain network and identify bottlenecks in the system.

Commercial trucking plays a major role in the transportation of goods throughout the supply chain. Thus, Economic Analysis is interested in calculating the reliability of the system being used by these private and for-hire fleets.

Efforts have been focused on the analysis of geo-positional data as it relates to tracking the inter-city, and to some extent intra-city transit times for commercial trucks. Distributions of transit times for selected city pairs in important trade corridors are analyzed with the intent to quantify the variability of the movement of goods.

This paper outlines the focus of the analysis thus far, preliminary results from the current methodology and directions for future

---

1 Views expressed in this paper benefited from exchanges between the authors and colleagues from Transport Canada. The authors thank all reviewers of this article for their useful comments. However, the views expressed herein do not necessarily reflect those of Transport Canada.

research. At a finer level, the paper provides a description of the data being used for the analysis, a description of the methodology used in manipulating the data, an overview of preliminary results for a subsection of the data and a discussion of the advantages and disadvantages of using purely geo-positional data for this type of analysis. The paper concludes with remarks details of the future direction of the research.

## The Data

Transport Canada has obtained data pertaining to commercial truck movements within Canada's national borders and a small portion of the United States spanning the January 2007 to May 2010 period from a private company that sells a positional tracking service to trucking fleets. The data consists of observations of longitude and latitude coordinates with corresponding date and time stamps for over 15,000 power units. This data is collected by a Global Positioning System (GPS) transmitter located inside the power unit. Most units are tractor trailer configurations.

The movement history of the unit typically transmitted to a third party database was then transferred to Transport Canada's databases after having been "scrubbed" to maintain the anonymity of the trucking fleets. The frequency of the location observations are on average every one-minute or every one mile travelled. The frequency of the data along with the number of power units being tracked resulted in a database with over 1.3 billion observations and was over 150 gigabytes in size.

## Methodology

The characteristics of the database and the focus of the analysis posed a significant challenge at the outset of the research. The volume of the data was large enough to restrict its manipulation using standard resources. In order to obtain results across the entire database, a Structured Query Language (SQL) application was developed to process the raw data into a structured format while also calculating initial estimates of average travel times. The data in its raw format did not contain information that would allow us to determine the beginning of a trip or the ending of a trip. Since the focus of the research was to measure trip specific variables, a methodology was developed in order to create pseudo trips based on transit time assumptions. This was later adapted to also include distance based assumptions.

The methodology required a perimeter or "Geo-fence" to be created for forty-eight city pairs of interest. Census subdivision boundaries were used for many of the cities. The process entailed mapping out the boundaries of the cities using Geographic Information System (GIS) software and inputting the geo-positional coordinates into the above mentioned application.

Using time-based logic, a movement was inferred to have taken place if a truck was found to be in a location of interest after being in a previous location of interest within a certain threshold amount of time. The transit time of the movement was calculated as the difference between the time the truck arrived in the destination geo-fenced area and the time it left the origin geo-fenced area. For example, if a power unit was located in the geo-fence comprising Montreal, and was later found in the geo-fence comprising Toronto within a maximum of 16 hours it would be deemed a trip.

In order to implement credible thresholds, a team of dispatchers was surveyed to determine the estimated average transit time and distance between selected city pairs. Maximum transit time thresholds that incorporated hours of service regulations for drivers were developed from these responses and the application used these estimates in its logic to determine trips. Maximum time thresholds were usually double the expected time of transit, in order to not exclude legitimate trips that may take longer to complete. The purpose of this processing was primarily to sort the raw data into a structured format comprising of trips to proceed to the next phase of analysis.

Once a trip was determined, it was grouped with other trips of the same origin and destination city pair, which resulted in distributions of transit times for the 41-month period of data. Estimates for the average transit time between city pairs, the 5th percentile time, the 95th percentile time, and the standard deviation were produced based on these results. The 5th and 95th percentile estimates were considered important based on consultation with experienced analysts with a background in transit time analysis.

The results were promising in that the average transit times were realistic, but the variability of even the shortest trips led to further adjustments to the process. The maximum time thresholds for the city pairs were sometimes too long, which resulted in non-direct trips and in some cases even multiple trips or multiple legs of a trip being included as one shipment. A further filter was needed to allow for the analysis of direct trips and the distance travelled by the truck would aid in this respect.

Distance was not a variable that was included in the original dataset; therefore it had to be calculated from the raw data. Implementing the haversine formula for calculating great circle distances[2], we were able to calculate the distance between each longitude and latitude coordinate. The total distance of an inferred trip was then calculated as the sum of the distances between each coordinate from the beginning to the end of the trip. This process was preferred to only taking the distance from the first coordinate and the last coordinate of the trip, because this would result in a straight-line distance between the two points and would not be accurate enough for the required filtering.

The fact that the data was being transmitted so frequently allowed for a fairly granular "bread-crumb" trail and a fairly accurate way of determining the distance travelled for a specific truck on a specific trip. City pairs were again granted maximum thresholds for allowable distance travelled on a trip and direct trips could now be inferred.

With a duration and distance calculation between each coordinate observation now in place, we could also measure the speed between

2 R.W. Sinnott, "Virtues of the Haversine", Sky and Telescope, vol. 68, no. 2, 1984, p. 159

each observation. We were also able to count the number of times speed was zero kilometers per hour for a particular trip, which indicated a stop and we could also measure the duration of time that the speed of the truck was zero.

This provided another method of filtering; one that allowed us to filter between express trips and casual trips by setting the maximum amount of time a truck was stationary. For example, when testing the stop time filter for the Vancouver to Toronto city pair, a maximum stop time threshold of 4 hours was used. It was assumed that if a truck was travelling from Vancouver to Toronto (a distance of 4,500 kilometers) and was not stationary for more than four hours at a time we could assume that it was not a casual movement, but most likely an expedited shipment.

**Table 1. Example of filtering parameters**

| | Measured in Hours | | | Measured in Kilometres | | |
|---|---|---|---|---|---|---|
| **Origin-Destination** | **Threshold** | **Expected** | **Max Stop** | **Avg Dist** | **Dist. L** | **Dist. U** |
| Montreal-Toronto | 16 | 8 | 1 | 545 | 518 | 572 |
| Edmonton-Calgary | 8 | 4 | 1 | 300 | 285 | 315 |
| Detroit-Toronto | 10 | 5 | 1 | 400 | 380 | 420 |
| Montreal-Quebec City | 6 | 3 | 1 | 250 | 238 | 263 |
| Regina-Saskatoon | 6 | 3 | 1 | 260 | 247 | 273 |

Table 1 presents an example of the filtering parameters that would be used to determine trips using the top five most frequent origin-city pairs, based on observations in the database. Edmonton to Calgary had an expected trip duration of 4 hours. The inter-city application used a buffer threshold of 8 hours for a truck to be in Calgary after leaving Edmonton to determine it as a trip. The distance filter would use 285 kilometers as the lower limit (Dist. L) and 315 kilometers as the upper limit (Dist. U) to filter non-direct trips out of the distribution. Finally, the maximum stop duration filter of one hour would eliminate trucks that were stationary along the trip for longer than that time.

As a result of this adapted methodology, two filters were now available to implement on the post-processed (GPS) data consisting of commercial truck activity. This provided distributions of transit times between city pairs of interest that could be further analyzed. The next section presents the results from each stage of this methodology.

## Results

Results from the analysis will be presented in the sequence of the above methodology and only for selected city pairs. The initial set of results from the post processed dataset will be discussed. Then, the results from the distance filtering will be discussed followed by the maximum stop time testing results.

From the raw database of 1.3 billion GPS observations, 468,915 trips were created by the application. The top 5 city pairs (not including return trips) with the largest number of inferred shipments were Montreal to Toronto, Edmonton to Calgary, Detroit to Toronto, Montreal to Quebec City and Regina to Saskatoon. See table 2 below for a depiction of the results.

**Table 2. Top five city-pairs by observation for the 1st stage of processing**

| | | Measured in Hours | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Origin-Destination** | **Trips** | **Mean** | **Std Dev** | **Min** | **Max** | **Expected** | **P5** | **P95** |
| Montreal-Toronto | 44,938 | 8.15 | 2.67 | 4.76 | 16.02 | 8 | 5.79 | 14.79 |
| Edmonton-Calgary | 33,492 | 3.89 | 0.81 | 2.54 | 8.00 | 4 | 3.15 | 5.65 |
| Detroit-Toronto | 29,752 | 5.49 | 1.62 | 3.03 | 10.01 | 5 | 3.76 | 9.11 |
| Montreal-Quebec City | 22,873 | 3.55 | 0.69 | 2.16 | 6.01 | 3 | 2.77 | 4.98 |
| Regina-Saskatoon | 17,588 | 3.33 | 0.56 | 2.10 | 6.01 | 3 | 2.68 | 4.44 |

Table 2 depicts the top five origin and destination city pairs, the number of inferred trips for that pair, the average transit time between city pairs, the standard deviation (Std Dev) of that transit time, the minimum (Min) amount of time and the maximum (Max) amount of time. Also shown is the expected number of hours from the dispatcher survey, as well as the $5^{th}$ and $95^{th}$ percentile times from each distribution of transit times.

Interpreting the Montreal to Toronto information shown in Table 2 we note that there were almost 45,000 observations from January

2007 to May 2010. The average transit time was 8.15 hours, while the fastest and slowest time that the trip was completed in was 4.76 hours and 16.02 hours, respectively. Figure 1 below depicts the distribution of the transit times for this city pair.

As can be seen by the figure it is not a normal distribution and is skewed to the right. The 5th percentile time was 5.79 hours. The 95th percentile time was 14.79 hours which, because of how long it seemed, caused some concern over the accuracy of this methodology.

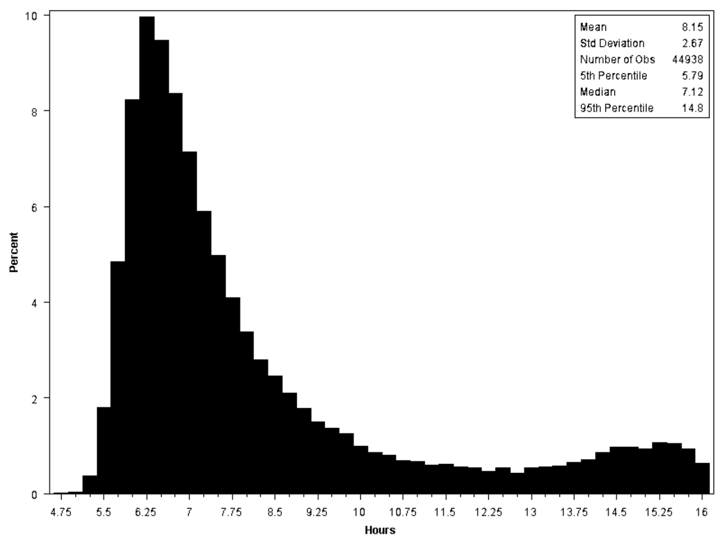


**Figure 1 Montreal to Toronto distribution of transit times**

To determine why there would be so much variability in the distribution some trips were selected to be analyzed inside of a GIS environment. Selected trips were plotted on a road network and analyzed. It was found that the time buffer we were allowing for a truck to reach its final destination was too large. In some cases, a truck was found to travel from the origin city to another city, back to the origin city and then on to the destination city of interest, all within the travel time threshold. Thus, three potential trips were being counted as one and this was skewing the results and implying more variability than was actually the case.

Figure 2 displays an example of two trips which, based on the initial time based assumptions, were included in the first-set of processed results. The figure details two distinct trips, trip #2473 and trip #13433, which left Montreal, QC and were then located in Toronto, ON within the maximum time threshold. Trip 2473 is shown to leave Montreal for Quebec City, return to Montreal and then head to its final destination in Toronto. Trip 13433 left Montreal and travelled through Ottawa to North Bay before heading to its final destination in Toronto.

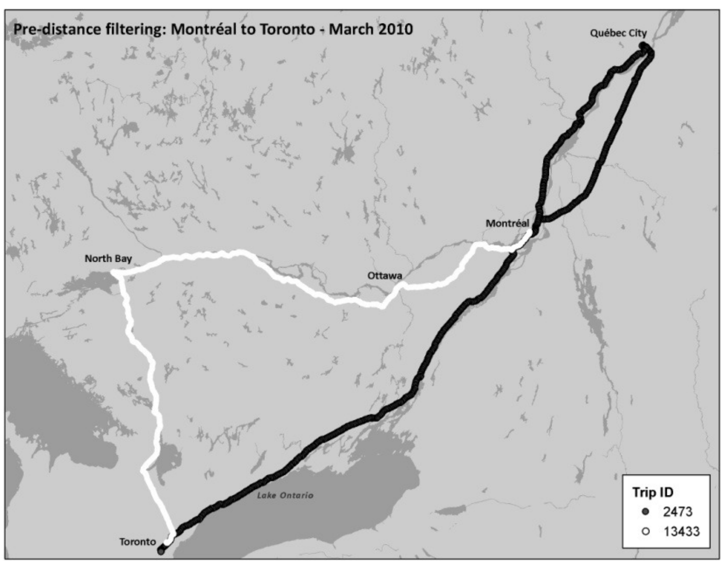


**Figure 2**

As mentioned above, these "erroneous trips" were products of the large buffer time allowed for a trip to take place. To solve this problem we needed to be able to determine the distance a truck was travelling for that trip in order to filter out the problem of detours and return trips being combined with single trips in the distribution.

Implementing the distance calculation on the entire database of observations was not feasible at the time. Analysis was done on an *ad hoc* basis when a city pair and time period of interest was required. Typically, a buffer of plus or minus five percent of the average distance for a selected city pair was allowed. The buffer was intended to account for different routings taken by a truck to reach its final destination.

A maximum stop time filter was also added to try to filter the data set into a more homogenous expedited movement subsample. See table 3 for the results of this process for the Vancouver to Calgary city pair for movements in the January 2009 to May 2010 time period[3]. The observed average distance for the dataset was 968 kilometers, implying a distance filter of +/- 50 kilometers. Given the distance of the trip, a maximum stop time filter of 2 hours was assumed to segregate expedited trips from the casual movements. The specific amount of time for this part of the filtering process which would produce the most representative subsample of the data is subject to future research. Using the above threshold of 2 hours removed 235 movements from this post-processed dataset for Vancouver to Calgary, leaving 601 trips to be analyzed.

**Table 3. Filter Results: Vancouver to Calgary**

| | | Measured in Hours | | | | | |
|---|---|---|---|---|---|---|---|
| **Dataset** | **Obs** | **Average** | **Perc5** | **Perc95** | **STD** | **Min** | **Max** |
| Raw | 836 | 14.83 | 11.94 | 19.98 | 2.31 | 10.99 | 20.97 |
| Distance filter | 746 | 14.62 | 11.93 | 19.55 | 2.19 | 10.99 | 20.97 |
| Distance and stop time filter | 601 | 13.86 | 11.81 | 16.44 | 1.39 | 10.99 | 19.76 |

With the two separate filters in place the average movement time decreases from 14.83 hours to 13.86 hours. The 5$^{th}$ percentile time is

3 The 17-month sub-sample of the Vancouver to Calgary data set was chosen instead of the 41-month dataset due to a focus on more recent trips and incorporating these results with analysis of other modal transit times of the same time period.

not affected very much as this represents the fastest time, or free-flow time, of the distribution and would already be at its near-minimum. The statistics that are most affected are the standard deviation and the 95$^{th}$ percentile. Both decreased as a result of trips involving detours being dropped, trips headed to other cities first, as well as trucks that may have stopped along the way to make deliveries or rests for longer than 2 hours. Thus, it is hoped that the distribution would now contain only direct and express trips and these would more accurately represent the typical movements under focus.

**Geo-positional data**

There are certain advantages and disadvantages to using geo-positional data for this type of analysis. The following section will detail some of these.

The majority of the time for the research thus far was spent dealing with the manipulation of the database. As mentioned previously, the data consists of 1.3 billion records containing only power id, date, time and latitude and longitude. Many hours were spent obtaining and processing the data into a useable format. Once Transport Canada began receiving the data and processing it correctly, the limitations of our resources became apparent. Upgrades were needed to house and process the data in a more efficient manner.

Another issue with the volume of the data was the processing time it took to produce the initial results with the in-house application. To explain this process, it would take approximately 90 minutes to generate trips and compute their associated times. The process utilized a SQL statement to isolate a month of data and convert it to a format that the application would recognize. It would take approximately 45 minutes to process each month and this resulted in roughly 35 million records.

The second step involved taking the formatted data and processing it with the in-house application which created the trips based on the geo-fences as well as the time parameters. This step also took roughly 45 minutes. The application would then output the trips and their corresponding latitude and longitude coordinates into a series of CSV (comma separated values) files based on trip origin and destination.

The distance formula was then applied to the individual CSV files to produce the estimated total distance, speed and stop times for each trip.

The process is lengthy and cumbersome. It is a product of the size of the database and the characteristics of the data set. With time, Economic Analysis has become more adept at manipulating the data and can now concentrate on improving the methodology to produce more accurate and enlightening results.

The size of the data was not the only problem that was encountered. Generating pseudo-trips proved to be difficult. Using time as the only parameter resulted in a large number of erroneous-assumed trips. To prevent this from happening it would be beneficial to have the purpose of the trip defined in the data stream. For example, a departure and arrival flag for a trip would be very useful in determining not only the starting and ending points of the trip, but intra-city movements could be identified as well.

It should also be noted that the characteristics of the truck were not detailed in the data stream. The size of the engine, the chassis type and other power unit specific information would be useful and if combined with the travel distance could be used to determine the annual fuel consumption or greenhouse gas emission of individual trucks or truck fleets as a whole, not to mention an indication of origin and destination pair or corridor footprint of such.

Also, the contents of the trailer were not disclosed so there was no way of knowing if the truck was carrying goods or not. Thus, a Truck-load, Less-than-load or empty designation would aid in the previous analysis of determining the transit time of loaded containers.

**Conclusions**

Insights on the distribution of transit times for commercial trucks were gained using the methodology cited above. Aggregation of individual transit times to produce the distributions needed to quantify variability is not straightforward. It was shown that there can be a large degree of variability in transit times if certain factors are not accounted for. This entails accounting for differences in the distance travelled for a truck or activity along the way such as if it performs multiple deliveries or one direct movement. We needed to be inclusive rather than exclusive when dealing with this data, meaning that we could not discard outliers in a trivial manner. The focus of the research is the variance of transit times and the issue of consistency is as important as reliability. Therefore, activity based filters were needed in order to transform the sample into a homogenous data set of units performing the same sort of shipments. Express shipments were preferred over multiple delivery shipments.

The methodology will be improved on and tested to incorporate more reasonable restrictions for time and distance. The GPS data used for this analysis has a very high transmission frequency and will be tested for its ability to determine intra-city movements in addition to the inter-city analysis described above. The hope will be to track movements from intermodal yards to determine high frequency destinations or corridors. This would include rail intermodal yards where trucks would pick up containers and transport them to a transload or distribution centre which we would consider the final destination for the goods.

Another potential area of research with this dataset is analyzing congestion points and areas that impact freight movements regularly. The ability to calculate average speed between observations enables us to locate road segments that have slow speeds when compared to others. Time of day analysis is also possible, though a re-tooling of the methodology will be needed in order to find common times of slow speeds amongst the thousands of trips.

## Bibliography

R.W. Sinnott, (1984), Virtues of the Haversine, Sky and Telescope, vol. 68, no. 2, 1984, p. 159