BUILDING FREIGHT SHIPMENTS FROM GPS DATA – FEASIBILITY STUDIES[1]

Serge Godbout, Statistics Canada Herbert Nkwimi Tchahou, Statistics Canada
Nathalie Hamel, Statistics Canada

1. Background

A review of the Canada Transportation Act (i.e. the Emerson Report; Canada, 2016) revealed that the federal government requires more and better quality transportation data to deal with many key issues, including commodity flows across the country. As part of this effort, Statistics Canada is redesigning its trucking statistics program, aiming to improve the survey coverage as well as simplify the sampling design and modernize data collection and processing. The evaluation of GPS data as a new source for trucking data has been identified as a priority by both Statistics Canada and Transport Canada.

A study has been underway since the summer of 2017 using GPS test files provided to Statistics Canada by Transport Canada (Legault and Veilleux, 2017; Broucke and Nkwimi Tchahou, 2018). These files contain anonymized GPS data from more than 40,000 trucks owned by 670 different companies and cover the time period between July and December 2016. The goal of this first study was to develop, test, and assess methods to translate the raw GPS data into concepts useful for the analysis of truck trips and the modeling of trucking freight movements.

2. Redesign of Trucking Statistics: A Primer

The objectives of the Trucking Statistics program are to measure the commodity movements and the activities of the Canadian trucking industry (Statistics Canada, 2018). The primary outputs are aggregate tables and a microdata file.

- The aggregate tables provide shipment statistics (counts, tonnage, distance, tonne-distance, revenues), with breakdowns by:
    - Domestic (origin and destination are in Canada) vs transborder shipments
    - Local (less than 25 km) vs long distance shipments
- The microdata file is shared with Transport Canada and used as input to the Canadian Freight Analysis Framework to provide a comprehensive picture of the freight flows across the country by geography, commodity, and transportation mode.

The program is about to be redesigned, targeting reference year 2019 to improve data quality and timeliness, reduce collection costs and streamline processing, and make full use of electronic reporting.

In order to expand population coverage, the transactional freight data source will add private trucking companies (i.e. businesses whose main activity is not trucking but who maintain a fleet of trucks to haul their own freight) to the for-hire trucking companies (i.e. businesses whose main activity is to undertake the transport of goods by truck for compensation). All selected companies will be asked to report their freight shipments (origin and destination, the commodity type, and the shipment tonnage and revenue) via an electronic file. Even though the respondent will be instructed to follow a prescribed format, it is expected that the collected data files will require some pre-grooming before being loaded into a standard format. Then, the shipment processing steps will include geography and commodity coding, calculation of derived variables (shipment distance and tonne-distance), imputation for nonresponse and estimation. The freight data will be combined with other survey data to improve the coverage and reduce the bias.

---

1 54th Annual Meetings of the Canadian Transportation Research Forum, May 26 - 29, 2019 at Vancouver, British Columbia

3. Introduction to GPS Data: Concepts and Notation

GPS data are generated by GPS devices connected to fleet management systems, which are increasingly used in the trucking industry. GPS data are available from the trucking company or through GPS service providers. As shown in Table 1, the core information included in the GPS data is minimal. Each record corresponds to a GPS signal called a ping, with the 3 categories of fields: Carrier and Vehicle ID (usually anonymized identifiers to prevent direct disclosure); date and time; and coordinates (latitude and longitude). The distances between coordinates are measured using the geodetic distance which follows great circles from the GEODIST function in SAS. Many researchers (Gingerich et al, 2016; Liao, 2014; Bernardin et al, 2014; Yang et al, 2014; Kuppam et al, 2013; Zhu et al, 2018) have successfully used GPS data in various studies, especially for trucking performance measures.

Table 1 – Example of a GPS dataset

Carrier ID	Vehicle ID	Datetime	Latitude	Longitude
A	A1	25/04/2018 8:03:23	45.555586	-125.896732
A	A1	25/04/2018 10:27: 44	45.235238	-125.714385
A	A1	25/04/2018 12:51:16	45.265958	-125.577085
A	A1	25/04/2018 15:15:52	45.445454	-125.642697
A	A1	25/04/2018 17:39:09	45.671174	-125.732463
A	A2	25/04/2018 10:31:20	47.109545	-132.401390
A	A2	25/04/2018 12:55:35	47.126438	-132.355647
A	A2	25/04/2018 15:19:58	47.244930	-132.813947

For our studies, test files were provided by Transport Canada, originally gathered by a GPS service provider from 670 different carriers. The data were collected from July to December 2016, recording 750 million pings from more than 40 thousand vehicles. The measured distance travelled amounts to approximately 2.18 billion km in total, which is similar to 14.7 Astronomic Units.

We performed an initial quality evaluation on the test GPS data. Coordinate rounding varies between either 6 or 4 decimals depending on the company, which correspond to an approximate precision of 0.15 or 15 meters respectively; this is not a major concern as the rounding is at most 10% of the estimated accuracy of 150 metres for GPS pings according to the literature (Gingerich et al, 2016; Liao, 2014). Similarly, the ping frequency is not constant but for the most part, the frequency is manageable with a median of ≤ 10 min for 80% of trucks. Rare inconsistencies were found, like unusual space-time jumps or middle-of-ocean locations, but we assume the dataset has likely been cleaned up, even if this was not confirmed. This test file is not meant to be representative of the target population of trucking companies but it contains a large amount of data to conduct various studies and proofs of concept.

4. Truck Behavior Analysis

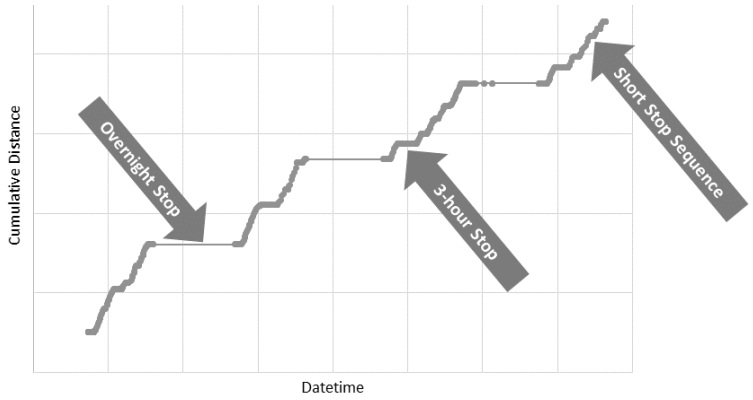
In order to analyze the behavior of vehicles, the identification of key events is necessary. The sequence of pings can be easily drawn using basic plots or specialized geomatics tools to visualize trips. When modeling shipments, the primary events of interest are trucking stops where commodities are picked up or dropped off. Figure 2 shows the cumulative distance of a truck over a week period. The slope of the curve is positive when the truck is in motion, so the plateaus represent periods of rest. In this diagram, there are overnight stops, long stops and a few sequences of short stops.

Trucking stops can be classified into the following categories:

- **Carrier yard**: Stops at the operating company;

- **Primary vs secondary**: Primary stops involve commodity transfer (pickup or drop off) while secondary stops are required to meet vehicle or driver needs or are imposed by the itinerary, like fuel fill up, coffee break, brake check, traffic jam, weighing station, cross-border customs, etc.

Figure 2 – Cumulative distance for a truck in a given week



Carrier yard stops are usually the most frequently made stops by trucks of the same company. They can be primary, especially in private trucking, or secondary, for shift changes or truck maintenance. Carrier yard stops are key to imputing the operating company of the truck, as shown in section 5.2.

Since no commodity transfers are implied, secondary stops are considered ignorable for commodity modeling. Gingerich et al. (2016) proposed the entropy score to identify secondary stops under the assumption they are visited by trucks from many different companies. From our study, we found this method may give false results in areas visited by a small number of companies from the GPS data files or for rarely visited stops. Also, we felt that some stops that are considered secondary for trucks from many companies should be classified as primary for trucks from other companies delivering to this location (e.g. gas stations). Instead of entropy, we preferred a criterion based on route deviation as secondary stops (carrier yard stops excluded) are expected to show negligible deviation when compared to their previous and next stops.

4.1 Identify Stops

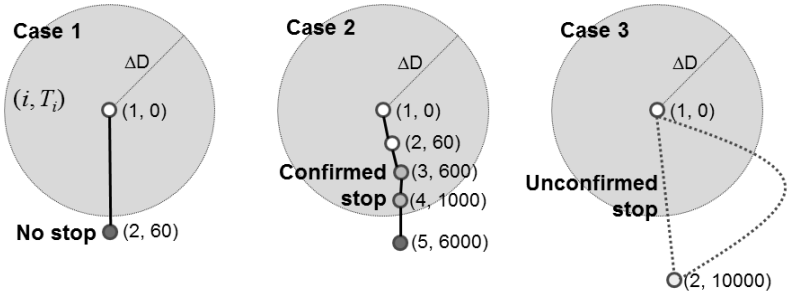
The naïve strategy to identify stops would be to look at null speed but this cannot always be observed given the ping frequency and the spatial errors. Also, the strategy needs to take into account the common behavior of trucks moving and having many short stops within the same location. An algorithm to derive trucking stops from a sequence of pings was proposed by Gingerich et al. (2016). A stop is confirmed if a truck has 2 or more consecutive pings inside a specified radius, within a minimum elapsed time. A stop is created from the set of the consecutive pings in a dwelling spell. From the examples shown in Figure 3 (where the time threshold is assumed to be between 60 and 600 seconds):

- Case 1 would not be a stop because the second ping is beyond the distance threshold set;
- Case 2 shows a stop from pings 1 to 4 since the 3rd ping meets the criterion and the 5th ping doesn't;
- Case 3 describes a censoring situation in which the time lag between consecutive pings is too large. This might be a stop (though location is unknown), but it is also possible there is a missing portion of the truck path (compare both dotted lines).

We tested many combinations of radius and time thresholds and the parameters proposed by Gingerich et al. (2016); a 250-metre radius and 15 minute interval based on spatial error analysis was initially shown to balance false positive and false negative errors, though further analysis has shown some missed trucking

stops. We now use less stringent parameters: a dwelling time of 3 minutes within a radius of 500 m. It is acknowledged this combination creates a large proportion of secondary stops, but we prefer addressing this at the shipment modeling phase rather than missing true stops.

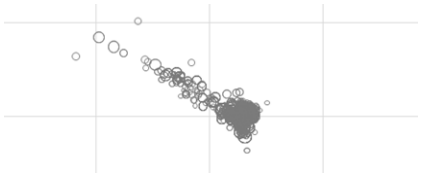
Figure 3 – Examples of stop creation from pings



4.2 Group Stops into Clusters

It is common that stops at the same location do not have the exact same coordinates, as shown in Figure 4. Clustering techniques are useful to avoid duplicating stops, identify loops in itineraries, and reduce the volume of data at processing. To build stop clusters, the initial strategy we implemented was to simply round the coordinates. Rounding to 0.015, which corresponds to an approximate precision of 2 km, gave good enough results for a quick implementation even though we acknowledge this method may combine distinct stops or split stop clusters in some cases. More powerful clustering techniques will be considered and evaluated later in the study.

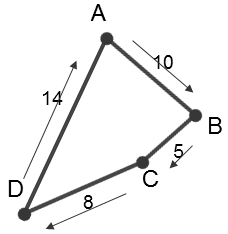
Figure 4 – Scatter plot from a cluster of a thousand stops



4.3 Build Truck Itineraries

A trucking itinerary becomes a sequence of visited stop clusters, usually creating a loop, especially at carrier yards, which can be represented as an oriented graph. Figure 5 shows an example of a simple trucking itinerary.

Figure 5 – Simple Truck Itinerary



- 5 nodes (A, B, C, D and A') representing the presence of a truck at a specific stop cluster from a given datetime period, where A' is the second occurrence of node A;
- 4 edges (AB, BC, CD and DA') from the pairs of 2 nodes the truck consecutively visited;

- 10 arcs (AB, AC, AD, AA', BC, BD, BA', CD, CA', DA'), from all the possible pairs of nodes, consecutive or not, ordered in time, representing origin and destination (O&D) combinations.

Distance and time measures are very important to consider when analyzing the itinerary of a truck, and it is common to use a distance matrix from graph theory. The elapsed time, the actual distance driven and the direct distance can be calculated for each of arcs. An important observation is that the actual distance driven is greater or equal to the direct distance and their difference is key when modeling shipment data. We assume that a shipment is more likely to happen if both distances are close; or, when actual distance driven is much larger than the direct distance, this is an indication that the end node is not the primary destination of the truck attached to this origin. From the itinerary in Figure 3 and the distances in Table 6, we can assume that the origin and destination (O&D) arcs AC, AD, AA', BA' and CA' are less likely to represent freight movements. Results would be even better if direct distances could be replaced by road network distances generated from open map sources, to avoid considering detours forced by the road network, such as getting around lakes, as criteria for stop classifications.

Table 6 – Distance Matrices from a Truck Itinerary

	Actual Distance Driven					**Direct Distance**					**Difference**				
	A	**B**	**C**	**D**	**A'**	**A**	**B**	**C**	**D**	**A'**	**A**	**B**	**C**	**D**	**A'**
A	0	10	15	23	37	0	10	11	14	0	0	0	4	9	37
B	0	0	5	13	27	0	0	5	12	10	0	0	0	1	17
C	0	0	0	8	22	0	0	0	8	11	0	0	0	0	11
D	0	0	0	0	14	0	0	0	0	14	0	0	0	0	0
A'	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

5. Freight Data Modeling

One of the main objectives of this study is to assess the feasibility of using GPS data as an alternate source for shipment data. The modeled shipments would be integrated into the trucking program as a planned replacement for data collection, as an alternate source for total nonresponse imputation, or to be used at estimation. For example, work on fusing trucking GPS data with survey information appears promising (Zhu et al, 2018). We tested linking stops to Statistics Canada's Business Register (BR), imputation of the operating company, and modeling of shipments.

5.1 Linkage of Stops to the Business Register

Statistics Canada's Business Register (BR) is a valuable source of detailed information on all businesses operating in Canada. In order to associate trucking stops to economic activities, we linked them to the operating entities on the BR by GPS coordinates. The stops were grouped by blocks larger than 500 m to take into account the accuracy of the GPS coordinates. This also significantly reduced the amount of records to be matched. All operating entities from businesses located within, or in the vicinity of, a block were linked to this block. For a specific stop, the operating entities linked to its block were scored and ranked using an index based on total revenue, main business activity, and geodetic distance to the stop. In general, the record linkage worked well but detailed examination of the results for some stops showed important challenges and significant limitations:

- Exact matching is difficult in high density areas or shared buildings. Similarly, there are very large locations exceeding matching limits;
- There are areas around stops without BR operating entities, like forestry or oil exploration, construction zones, rest areas, traffic jams, or operating entities located in the US;
- There is a significant gap between the geographic location concepts. The trucking stops point to parking lots while the BR coordinates come from civic addresses, located on the streets or the street blocks.

Consequently, proximity was not the only criterion used to match stops to companies from the BR. We found many cases where the most likely business attached to a stop does not necessarily have the address closest to the stop. We revised the linkage objective from matching to a single business to a more flexible process that involves building an inventory of the businesses in close proximity of each stop. This allows assessing the expected economic activity and how they relate to the other trucking stops from the itinerary.

5.2 Imputing Operating Companies

The imputation of the operating company is required for the integration of the modeled shipment into the Trucking Statistics program. Also, information on the operating company from the BR, such as its main business activity, total revenue, or number of employees, will help associate the commodities to the modeled freight movements. The supporting assumption is that the truck is owned by or operates for the most frequently-visited business in its stop network. From the results of the linkage to the BR, a score by company was calculated from the stop dwelling time, the industry classification, the allocated revenues from the BR operating entities, and the calculated distance between the coordinates from the BR operating addresses and the stops. The company with the highest total score was imputed as the operating company.

5.3 Build Shipment Data

In order to model freight movement, the critical step is to identify, from the set of all the nodes on the itinerary of a truck, which ones are the most likely to correspond to commodity pickup and drop off, so that their corresponding O&D arcs would become the shipments.

As an initial strategy to identify shipments, we tried to predict the main commodity flows between the nodes of an itinerary. A mapping table between industries (NAICS) and commodities (NAPCS) was built from the Annual Survey of Manufacture and Logging and the Annual Wholesale Survey data. This way, we were able to allocate the revenues of the businesses in close proximity of each stop down to input and output commodities. Commodities repeated through the itinerary were used to select and weight the O&D arcs of interest. An optimisation algorithm was applied to maximize the freight movements under some constraints. The mechanics worked fine but the optimisation results showed that this strategy puts too much emphasis on commodity flows and not enough on the observed behaviour of the truck.

The new strategy to identify the shipment data will start from the truck itinerary. All possible O&D arcs will be listed, with auxiliary variables attached to predict the likelihood of a freight movement using predictive methods, including machine learning. Here is a list of potential predictors:

- **O&D Arc**: Distances (actual driven, direct, and difference), elapsed time, average speed, frequency in truck/carrier's history, links to other shipment sources (e.g. trucking data), etc.
- **O&D Nodes**: Frequency in truck/carrier's history, elapsed time, presence of businesses in the close area, etc.
- **Itinerary**: Actual distance driven, elapsed time, average speed, number of nodes and arcs, etc.
- **Operating company**: Main business activity (NAICS), carrier yard and terminal locations, etc.

From the estimated likelihood of each arc, we will run an optimization algorithm to select the O&D arcs to maximize the yield of the truck's itinerary under constraints on maximum load capacity (100%) and trip types. In the final stretch, modeled shipment records will be created from the selected O&D arcs and the missing information will be added using information from trucking data or other sources.

- **Commodity**: Coded from the auxiliary variables from the selected O&D arcs;
- **Shipment tonnage**: Imputed using average values;
- **Shipment value**: Imputed using an average ratio per ton;
- **Shipment tonne-distance**: Derived from imputed tonnage and calculated actual distance;

- **Shipment revenue**: Imputed as an average ratio per ton-distance.

Imputed shipments can be adjusted to control totals reported by companies through calibration, benchmarking, or other techniques.

6. Empirical Study

As a proof of concept, we processed the first 200 out of 670 carriers. The stops were derived, the most frequent stop clusters were matched to the BR, and operating companies were identified using the score explained in Section 5.2. Table 7 provides a summary of the results, including 8 carriers that could not be matched since there were no businesses from the BR in the vicinity of the most frequent stop clusters (e.g. US companies). The remaining 192 carriers were split into 3 tier groups based on their matching score. The tier 1 had a matching rate of 73% (47/64) while less than 50% of the revisions done to tier 2 carriers confirmed the match. The matching results of the tier 3 carriers were not reviewed.

Most of the matching errors found at review were due to inaccuracy of the coordinates derived from the BR addresses or profiling issues. At the end of this study, the automated linkage and the manual review were able to confirm the links of 99 carriers to operating companies: 71 from general freight, 18 from specialized freight, 5 from private trucking, 1 from construction, and 4 not listed on the BR. We are confident that the matching rate would improve with better resources and tools.

Table 7 – Summary of Matching Carriers to BR companies

Matching Score	Total	Reviewed: Confirmed	Reviewed: Edited	Reviewed: No conclusions	Not Reviewed
Tier 1 – High score	64	47	12	5	---
Tier 2 – Medium score	64	16	23	7	18
Tier 3 – Low score	64	---	1	---	63
No possible match	8	---	---	---	8
Total	**200**	**63**	**36**	**12**	**89**

We compared the shipment data from the existing Trucking Commodity Origin and Destination survey (TCOD) with the 99 operating companies from the GPS data and found that 25 companies were in common with our 6-month reference period. The shipment data were extracted from the TCOD database for 24 of these 25 companies; one company was ignored because of volume issues. From the initial set of 4,213 shipments from the 24 companies in TCOD, 1,864 records which had US origins, US destinations, or invalid datetime values were dropped, leaving 2,349 shipments for the study.

Table 8 – Distribution of distance and datetime gaps between TCOD shipment data and GPS modeled shipments (24 companies, July-December 2016)

Distance Gap	Nb Shipments	%
0-2 km	874	37%
2-5 km	798	34%
5-10 km	392	17%
10-25 km	171	7%
25+ km	114	5%
Total	**2,349**	**100%**

Datetime Gap	Nb Shipments	%
0-5 days	472	20%
5-25 days	618	26%
25-50 days	437	19%
50-100 days	493	21%
100+ days	329	14%
Total	**2,349**	**100%**

Since TCOD coded origin and destination using postal codes or ZIP codes, coordinates were imputed using the average lat/lon by postal code calculated from operating entities on the BR. Each of the 2,349 shipments from TCOD was matched to the GPS data from the same company by lat/lon, choosing the

O&D arc having the smallest maximum distance between origin or destination coordinates. The distribution of the distance gaps is given in Table 8. We observed that the accuracy looks very good since more than 70% of shipments had a distance gap smaller than 5 km, despite the error sources listed before (precision of GPS data, concept gaps between GPS and BR coordinates, imputation of coordinates from postal codes). The distribution of the datetime gap, calculated as the maximum absolute difference between the shipment data from TCOD and the start and end dates from the matching GPS arcs, looked less accurate than expected because of the lower quality of the datetime field from the TCOD data and consequently datetime was not used when matching TCOD shipments and GPS data.

7. Conclusion and Future Work

Algorithms tested to derive and classify stops and to associate GPS data to companies have shown to be powerful and easy to implement. On the other hand, there remains a lot of work to do on freight modeling and parameter optimization; in particular, the new freight movement modeling strategy focusing on truck itineraries has to be tested and evaluated. Also, enhanced distance calculations using road network and data visualization from powerful Geomatics techniques should ultimately improve the accuracy of results and their analysis.

Besides dealing with large volume data and limited resources, the most important challenge we faced was the conceptual gap when matching the GPS stops to the operating entities from the Business Register. We strongly encourage adding a more flexible Geography Information System (GIS) into the process model to bridge this gap. To get ready for using GPS data on a production scale, we also recommend working with partners for the continuous acquisition of GPS data and for the development of processing tools for GPS data. Beyond these considerations, collaboration with external partners like Transport Canada, provincial departments, American agencies and academic researchers is key to success.

8. References

Canada (2016). Pathways: Connecting Canada's Transportation System to the World. Transport Canada: Canadian Transportation Act Review.

Bernardin, V.L. Jr., Trevino, S. and Short, J. (2014). Expanding Truck GPS-Based Passive Origin-Destination Data in Iowa and Tennessee. Transportation Research Board 2014, Annual Meeting.

Broucke, H. and Nkwimi Tchahou, H. (2018). Modélisation des mouvements de marchandises par camion à partir de données GPS. Feasibility Study Phase 2. Technical Report. Business Survey Methods Division, Statistics Canada.

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

Kuppam, A., Lemp, J., Beagan, D., Livshits, V., Vallabhaneni, L. and Nippani, S. (2013). Development of a Tour-Based Truck Travel Demand Model using Truck GPS Data. Transportation Research Board 2014, Annual Meeting.

Legault, J. and Veilleux, L. (2017). GPS Data. Feasibility Study Phase 1. Technical Report. Business Survey Methods Division, Statistics Canada.

Liao, C.-F. (2014). Using Truck GPS Data for Freight Performance Analysis in the Twin Cities Metro Area. Research Project. Final Report 2014-14. Minnesota Department of Transportation.

Statistics Canada (2018). Transportation Data and Information Hub. https://www144.statcan.gc.ca/tdih-cdit/index-eng.htm.

Yang, X., Sun, Z., Ban, X. and Holguín-Veras, J. (2014). Urban Freight Delivery Stop Identification Using GPS Data. Transportation Research Board 2014, Annual Meeting.

Zhu, S., Amirjamshidi, G. and Roorda, M. J. (2018). Data fusion of commercial vehicle GPS and roadside intercept survey data. *Transportation Research Record*, **2099** (1), p. 102-112.