# MINING FREIGHT TRANSPORTATION ACTIVITY LOCATION & TYPE USING LARGE-SCALE PASSIVE GPS DATASETS

Kevin Gingerich, Hanna Maoh, and Bill Anderson
University of Windsor

## Introduction

As the mass production of goods continues to grow and rapid technological advances take place, modern cities have become increasingly specialized. This agglomeration has long been known to provide benefits that increase production efficiency, giving firms an advantage in a globalized market (Marshall, 1890). A result of this agglomeration and the globalized marketplace is that trade becomes more prevalent with specialized economies trading amongst each other. For example, truck freight in Canada doubled in tonnage between 1992 and 2004 (HRSDC, 2013) and nearly doubled in value in the US between 1993 and 2002 (USDOT, 2013). This increased economic and transportation dependency underlies the need to study regional freight transportation patterns.

Regardless of the current trends in freight models, the need for input data remains a consistent issue. Previous data was often obtained through sampled surveys that can be expensive for the modeller and laborious for the respondent. An emerging alternative to surveyed data is to obtain passive information from technology such as global positioning systems (GPS) that are frequently used by trucking firms to track the current position of their fleet. While privacy concerns may limit the availability of such data for public research interests, it provides enormous potential in understanding the current patterns of freight based on the large volume of data alone. But this also presents an enormous challenge since the data is not originally created for this

purpose. Therefore a need exists for methods to efficiently generate useful findings.

This paper introduces a number of techniques to mine a very large set of GPS data points that pertain to truck movements in Canada and the U.S. The data were obtained from Transport Canada and contains individual GPS pings for a group of freight carriers. While the pings occur throughout Canada and the U.S., these carriers are all Canadian owned. Although the full dataset covers several years of entries, this paper uses a subset containing all GPS records for the month of March, 2013. During this time a total of 750 unique carriers encompassing 40,650 individual trucks and approximately 101.6 million individual GPS pings were recorded. Each GPS ping results in a data record containing the following information: carrier ID; power (truck) ID; latitude; longitude; date; and time.

Two objectives are discussed in this paper. The first research objective is to develop a computationally simple approach to cluster GPS pings together and determine the location of carrier shipping depots. A high density resulting from these depot locations provides a strong indication of the regions that generate a large amount of freight activity. From a practical perspective, these areas are of interest to us since they will comprise a large proportion of trip activity production in our sample compared to areas with no shipping depots. Moreover, the location of these depots provides a location to delineate truck tours beginning and ending at their shipping yard.

The second part of this paper explores a method for differentiating between two types of stops:

1. Primary stops: a transfer of goods takes place between the truck and location (or another truck)
2. Secondary stops: the truck is stationary for other purposes such as driver breaks or fuel refills

Since travel diaries do not accompany the GPS data we are utilizing, the purpose of any identified stops based on dwell-time is initially unknown. However, we expect that secondary stop locations such as truck stops tend to attract a variety of trucks and carriers compared to

primary stop destinations. This heterogeneity in the carrier types visiting and dwelling at a stop location can be captured by utilizing the concept of entropy. The latter can be used to identify the level of heterogeneity in the trucks gravitating to a particular location. To this end, the second research objective of this paper is to apply and examine the concept of entropy to the GPS data to help differentiate primary stops from secondary stops. This is an important distinction for freight models since the former stops are used to identify trip ends for a given truck. To our knowledge, this is a novel approach to understanding the purpose of a stopped truck that is made possible by the large amount of GPS pings observed in the dataset.

The remainder of this manuscript begins with an overview of passive GPS data and its use in freight literature. This is followed by an analysis of our two research objectives: (1) identifying carrier shipping yards and (2) identifying stops occurring for secondary purposes. Finally, the limitations of these techniques and future applications are discussed in the conclusions.

**Literature on GPS data and Freight Models**

A prominent trend in modern spatial geography is the utilization of data voluntarily created by the general population labeled by Goodchild (2007) as volunteered geographic information. While freight models using global positioning system (GPS) data have been limited in the past, it is emerging as a popular data source for passenger and freight information due its ability to triangulate the position of a device in a timely manner, reduced respondent burden for participants due to passive collection, and its recent integration into many devices such as phones and cameras.

For GPS data obtained from truck freight, many of the same issues persist compared to passenger travel though the data and information may be harder to obtain for confidentiality purposes. Some of the recent uses of GPS data for freight modelling include identifying trip ends (Du and Aultman-Hall, 2007), trip end clustering (Sharman and Roorda, 2011), travel time reliability near major ports (Chu, 2011), border crossing times and variability (Leore et al., 2003) time

intervals between freight deliveries (Sharman and Roorda, 2013), and vehicle congestion and reliability on road segments (Zhao et al., 2013). In addition, map matching the GPS data onto a road network for GIS applications frequently performed to convert the point data to digital road links (Shawathe, 2007; Zhao et al., 2013).

While GPS data can provide an unobtrusive method of data collection, the issue of accuracy needs to be addressed before it can be relied upon without the addition of other data sources. Bohte and Maat's (2009) initial validation of a study identifying trip purpose and mode from GPS passenger data had an overall accuracy of 43% and 70%, respectively. Gong et al. (2012) measured the accuracy of their mode choice algorithm of GPS data of pedestrians in New York City with an overall success rate of 82.6%. Du and Aultman-Hall (2007) measured the validation results of their model identifying trip ends at 94%. Therefore a high amount of accuracy is possible with GPS data alone but this will vary by the approach used, sample size, spatial coverage, and the accuracy of the GPS receiver. In the latter case it is helpful if each ping has a dilution of precision (DOP) measurement to provide an indication of signal error.

In addition to the applications of freight data outlined above, modelling truck tours is another important topic for understanding freight movements. A truck tour typically starts and ends from some form of base depot such as a truck yard or a distribution centre (Ruan et al., 2012). Unfortunately passive data such as GPS pings do not generally denote these locations without additional truck diary information. Kuppam et al. (2014) identified a truck tour start point as the first starting point at the beginning of a day for a given truck. While this method works for urban models, longer distance inter-regional truck tours cannot be differentiated using the date. In such a case, identifying the truck yard for a carrier provides a known location to delineate truck tours. Sharman and Roorda (2010) briefly discuss identifying truck yards using the number of visits by trucks and the length of stop durations but this was not yet implemented. Our paper formulates an approach to identifying these shipping depots that can be efficiently performed with a large quantity of data.

Another important topic for freight transportation modelling is identifying a primary stop where a transfer of goods takes place. These events are significant when observing truck movements because they represent the start/ends of individual truck trips. Moreover the primary stop locations can be used to identify the type of goods that are shipped by a given truck. One method used to differentiate the type of stop is by spatial proximity to known locations. For example, Bohte and Maat, (2009) used known locations of truck stops and gas stations to identify secondary stops. But a comprehensive list of all secondary stop locations becomes increasingly difficult as the size of the study area and the number of political boundaries increases. Alternatively, there are other spatial variables that can be employed to determine if a stop was used to transfer goods including dwell time (Du and Aultman-Hall, 2007; Gong et al., 2012), trajectory change, and distance to a major road (Du and Aultman-Hall, 2007). We believe that the variety of truck carriers observed at a location can also be used to help identify the purpose of a stop but this has not been previously studied to the best of our knowledge.

**Identifying Shipping Depot Locations**

The first research goal of this paper is to develop an efficient method of determining the location of base depots utilized by trucks. An aggregation of these depot locations by municipality or other political boundaries is a useful proxy for the regions with high levels of trade activity originating from them. Moreover, their locations provide a convenient location to delineate truck tours where a truck begins and ends at their carrier's main yard.

Identifying the location of a carrier's main shipping depot was performed on the basis of the following two assumptions:

1. The depot will occur at a location that is frequently visited by trucks belonging to the given carrier
2. The first GPS ping obtained from each truck is more likely to occur at the shipping depot compared to later GPS pings

The first assumption reflects the frequency of trucks stored at a truck yard overnight or for longer periods of time between truck tours. Moreover, these yards can sometimes include the transfer of goods on site. But there are other locations that can also be frequented by trucks such as major highway interchanges and truck stops. A false positive could erroneously identify these other locations instead of the shipping depot if all GPS pings are utilized. The second assumption was included to prevent such a scenario from occurring. We expect that the first observed GPS ping is the most likely point to occur at the shipping depot as the vehicle starts towards its intended destinations. Of course, the first point is not always going to correspond to the shipping depot. For example, truncation of the data at any given date results in some of the first pings occurring in the middle of a trip. But if the first point from each truck has a moderate probability of occurring at a shipping depot, a density surface from the first point of all trucks should reveal the true location.

This problem in its simplest form corresponds to the calculation of densities that would often be best suited for GIS based programs such as ArcGIS. But the runtime using such software increases significantly due to the size of the dataset, size of the study area stretching across Canada and the U.S., and the required precision for a final location. Therefore an alternative was developed that could be quickly processed in database software such as MS SQL to reduce the overall runtime and simplify the entire process. This method uses the following procedure:

1. For a given carrier *c*, calculate a single coordinate $Z_{ct}$ for the first GPS ping of each truck *t* as follows:

$$Z_{ct} = X_{ct1}Y_{ct1} \quad (1)$$

Where $X_{ct1}$ and $Y_{ct1}$ are the longitude and latitude coordinates from the first GPS ping recorded for truck *t.* Without combining the coordinates into one value ($Z_{ct}$), the latitude or longitude coordinates can occur frequently along roads facing a cardinal direction. This phenomenon negatively affects the final results.

2. Next, the coordinate $Z_{ct}$ is rounded to $p$ significant digits. The rounding procedure is used to group nearby points together. A lower value of precision (*p)* results in a larger area that is used to bind the points into a cluster. The resulting coordinate ($Z'_{ct}$) is used as an identifier for the points clustered together.

$$Z'_{ct} = ROUND(Z_{ct}, p) \tag{2}$$

Where $Z'_{ct}$ is the coordinate value of $Z_{ct}$ rounded to $p$ significant digits.

3. The cluster with the largest number of points for a given carrier is determined by taking the mode of $Z'_{ct}$ for all trucks *t* belonging to carrier *c*.

$$Z'_{c(depot)} = MODE\{Z'_{ct}\} \tag{3}$$

Where $Z'_{c(depot)}$ is the identifying coordinate for the largest cluster of points belonging to carrier *c*.

4. Finally, the longitude and latitude coordinates for the carrier's shipping depot are calculated as the average coordinates from all points that have the $Z'_{c(depot)}$ coordinate.

$$X_{c(depot)} = \frac{\sum_t^n X_{ct1}}{n} \; ; \; X_{ct1} \in \{Z'_{c(depot)}\} \tag{4}$$

$$Y_{c(depot)} = \frac{\sum_t^n Y_{ct1}}{n} \; ; \; Y_{ct1} \in \{Z'_{c(depot)}\} \tag{5}$$

Where $X_{c(depot)}$ and $Y_{c(depot)}$ are the longitude and latitude coordinates of the shipping depot for carrier *c*, $X_{ct1}$ and $Y_{ct1}$ are the coordinates for the first GPS ping observed for truck *t* belonging to carrier *c*, and *n* is the total number of trucks belonging to carrier *c*.

The result of this algorithm is a table of coordinates corresponding to the primary shipping depot for each carrier. Note that if the largest cluster of points has less than three GPS pings, the algorithm provides an output of 'null' to avoid misleading results. This means that there

is no cluster large enough to identify the carrier yard. This becomes more common when a given carrier has fewer trucks in the sample. From our GPS dataset for March, 2013 there were 472 out of 750 carriers that had an identifiable yard.

The results from the coordinate mode procedure were validated using kernel density estimation in ArcGIS to ensure that our algorithm provides a proxy for the densest location. Comparing results between the two methods conducted on GPS points for the 472 carriers, there were 50 outliers. A manual verification of all these points revealed that this was due to the occurrence of multiple locations with a dense cluster of points. Accordingly, future uses of this technique could be adjusted to consider the possibility of multiple shipping yards. Removing the 50 outliers from the results, a statistical analysis of the two methods revealed a root mean square error (RMSE) of 70 meters. The remaining error is partially driven by the resolution of grid cells in the raster oriented GIS density approach but nonetheless indicates that the SQL algorithm is performing correctly. The runtime for the process using ArcGIS lasted approximately 6 days although periodic interruptions would occur that stalled the program until user intervention occurred. The algorithm using the process described in this paper was implemented in MS SQL Server and completed in roughly 3.5 hours on the same computer resulting in a considerably faster alternative.

A second validation was performed to test whether the results correspond to the spatial location of a carrier shipping yard. An in-house software tool was developed to identify the type of land use through two spatial databases (Google and Factual). Using this tool on a 10% random sample (50 points) revealed that 100% of the locations resulting from our methodology are located at a truck yard.

**Determining Truck Stops using Entropy**

The second technique discussed in this paper uses the clustering approach developed earlier in this paper to study locations where trucks have stopped. The full set of stops is composed of any point where a truck remains relatively stationary for 15 minutes or longer.

The purpose for each stop can be aggregated into two categories. Primary stops occur when a transfer of goods is taking place to load or unload goods to a truck. Secondary stops occur for other reasons such as fuel refills or driver breaks. To differentiate between stop types, we applied theoretical concepts of entropy to the application of stopped trucks.

Entropy is a well-known principle describing the chaos / dispersion of a system that is used in many different disciplines. For example, it can be used in land use planning to describe the heterogeneity of land uses with a larger value associated with a greater mix of these land use types. In this case, an Entropy Index *(EI)* is created to quantify the variety of carrier fleets using a particular stop location. More carriers visiting a particular stop will result in a larger *EI* value. Conversely, a stop with only one corresponding carrier will have an entropy value of zero as shown in Figure 1. We expect primary stops where the loading/unloading of goods takes place to exhibit fewer carriers associated with the stop compared to secondary stop locations. Therefore the stops with a higher *EI* (and a higher number of carriers) will be more likely to provide a secondary function compared to stops with a lower *EI* (and fewer carriers). If this expectation holds true, we can tease out records from the full list of stops to isolate for primary stops where goods are transferred.

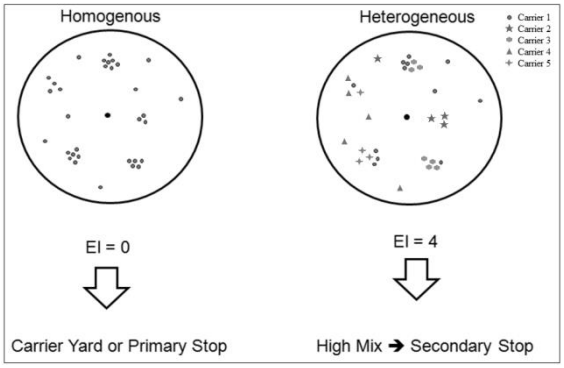


**Figure 1: Entropy index interpretations**

The entropy index (*EI*) for each cluster is calculated using the following formula:

$$EI = -\sum_{c=1}^{C} \frac{N_c}{N} \; x \; Ln \left(\frac{N_c}{N}\right) \tag{6}$$

Where $N_c$ is the number of stops occurring for a given carrier $c$ in the cluster, $N$ is the total number of stops in the cluster for all carriers, and $C$ is the total number of carriers.

A histogram of the results shown in Figure 2 for the one month dataset exhibits a range from 0 (only one carrier) to 4.78 for 3372 cluster locations where trucks stopped. An interesting outlier in the histogram is the large number of clusters with an *EI* value between 0.6 and 0.8. An investigation of this phenomenon showed that a value of roughly 0.69 occurs frequently due to the presence of two carriers with relatively equal proportions of trucks stopping at a given location. A proportion of exactly 50% for both carriers results in an *EI* value of 0.693. This is the maximum value of *EI* that can be obtained for a scenario involving two carriers.

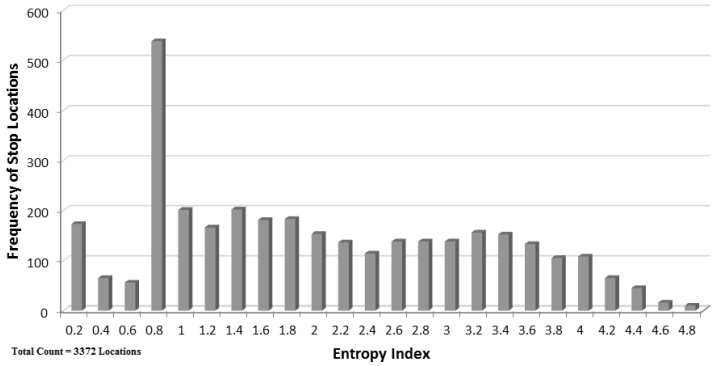


**Figure 2: Histogram of entropy results**

The original hypothesis for the entropy index was that a high entropy would correspond to stops used for secondary purposes such as rest stations. To validate this result, the top 150 entropy index values were manually checked to determine the type of stop corresponding to the clusters (*EI* values ranging from 3.97 to 4.78). From these 150 location clusters, 148 of them correspond to secondary stops such as truck stops, gas stations, and several motels. The last remaining two locations are possible transfer points for goods since they were located in parking lots for big box retailers. This initial validation establishes that the original hypothesis holds true.

Another validation was conducted to determine a value of *EI* where the majority of cluster located above the threshold value correspond to rest stops and locations with *EI* values below the threshold are generally not. This value is useful to apply in freight models as a rule when establishing the purpose of a stop. A stratified random sample was conducted to determine if the cluster was located at a rest stop or not. The first sample consisted of 250 random clusters and demonstrated that values above an *EI* of 3 were mostly rest stops and values below an *EI* of 2.5 did not generally correspond to rest stops. A subsequent random sample with 150 random clusters with *EI* values between 2.5 and 3 was used to determine a precise value that separates the majority of primary and secondary stops. Subsequently, the total sample of 400 points (10% of the total number of clusters) was used to determine a threshold value of 2.75. As shown in Figure 3, 90% of the cluster locations above the threshold value of 2.75 were found to belong to rest stops. Conversely, only 38% of the cluster locations with *EI* values below the threshold value belong to rest stops.

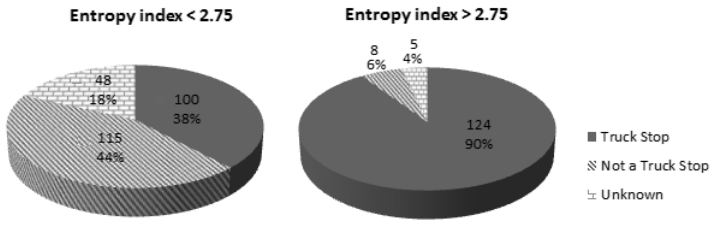


**Figure 3: Validation results for locations with entropy index values less than 2.75 (left) and greater than 2.75 (right)**

In summary, locations exhibiting a large degree of heterogeneity among carriers with trucks stopping at the location are likely occur for secondary purposes. Locations with an entropy value greater than 2.75 are 90% likely to occur for secondary purposes. Therefore the entropy technique developed here can be used to improve the accuracy of establishing the stop purpose in freight models when this is not provided in the original data.

## Conclusions

The methods discussed in this paper use passively collected GPS pings for truck movements belonging to Canadian carriers. The large size of the dataset provides a substantial challenge but also provides opportunities that are not available when using smaller datasets. Two research objectives are explored in this paper to enhance the capabilities of freight transportation models. These methods are applied to a dataset covering one month of truck movement consisting of 101.6 million individual GPS pings for 40,650 trucks.

The first objective is to create a computationally efficient process to determine the location of shipping depots. An approach is developed to cluster the first GPS ping from every truck for a given carrier. The location of the cluster with the largest number of points corresponds to a shipping depot for the given carrier. Validation of the results confirms that this method reveals the same locations as more time consuming density calculations performed in spatially driven software. Moreover, a 10% sample of these points was validated

manually using aerial photography and each location coincided with a shipping depot.

The locations of these shipping yards are significant for several reasons. They provide insight into the regions that have a high level of trade activity. If resources are limited it would be beneficial to focus on these regions given their economic importance in terms of freight movement. These shipping depots also provide a convenient location to delineate the start/end of truck tours. In the future, the shipping yard location technique could be expanded to allow for the possibility of multiple sites per carrier. The current version of the technique only identifies a single location corresponding to the highest density of points but an analysis of the data shows that multiple shipping depots are possible for large carriers.

We also used the clustering procedure described in this paper to create spatial groups of points representing stopped trucks. An entropy index was formulated and applied to each cluster for analysis. A larger entropy value indicates that a greater degree of heterogeneity exists among carriers stopping at a given location. A validation of the results confirms that stops with a larger entropy index are more likely to fulfill secondary needs such as gas refuelling and rest breaks. Locations with an entropy index greater than 2.75 have a 90% chance of corresponding to a secondary stop. This is true for only 38% of the locations with an *EI* below 2.75. However, this threshold may not be transferable for other datasets since the entropy index is dependent upon the amount of input data and the cluster size.

The clustering algorithm formulated here provides a simple approach clustering GPS points representing stops. However, there is the potential for a cluster of points and other outlier points to have the same location coordinate ($Z'_{ct}$). While constraints are in place to disregard the outlier points from any calculations, the result of this limitation is that each point may not have a final value attached to it. For the location of carrier yards, this has no negative implications since we are only looking for the largest cluster of points. The clustering procedure also worked well to analyze groups of stopped trucks by quantifying the entropy resulting from multiple truck

carriers. However, implementation of the entropy technique in a truck movement model would generally require each stop to have an entropy value. In this situation, the entropy could be calculated differently at the cost of a longer processing time - individually for each point or by using a more advanced clustering method.

The techniques introduced here represent a novel approach to extracting information from a very large passive GPS dataset that the authors have not seen in previous literature. The final outcome of the methods presented here is that we can delineate many of the truck tours and associate an industry with them after identifying the locations of goods transfers. In the future, this work will enable us to better understand the truck freight moving across important trade corridors by separating the data by region of origin/destination and industry.

**Bibliography**

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. *Transp. Res. Part C*, 17, 285-297

Chu, H-C. (2011), An empirical study to determine freight travel time at a major port. *Transp. Planning and Tech.*, 34(3), 277-295

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. *Transp. Res. Part A*, 41(3), 220-232

Gong, H., Chen, C., Bialostozky, E. and Lawson, C.T. (2012) A GPS/GIS method for travel mode detection in New York City. *Computers, Environment and Urban Systems*, 36(2), 131-139

Goodchild, M.F. (2007), Citizens as sensors: The world of volunteered geography. *GeoJournal*, 69(4), 211-221

Human Resources and Skills Development Canada (2013), Accessed on Oct. 22, 2013, <http://www4.hrsdc.gc.ca/.3ndic.1t.4r@-eng.jsp?iid=67#M_5>

Kuppam, A., Lemp, J., Beagan, D., Livshits, V., Vallabhaneni, L. and Nippani, S. (2014), Development of a tour-based truck travel demand model using truck GPS data. TRB 2014 Annual Meeting Proceedings

Leore, B., Trent, M. and Shallow, T. (2003), Using truck tractor logs to estimate travel time at Canada-U.S. border crossings in Southern Ontario. *Proceedings of the 38th Annual CTRF Conference*

Marshall, A. 1890, *Principles of economics*. London: Macmillan.

Sharman, B.W. and Roorda, M.J. (2013), Multilevel modelling of commercial vehicle inter-arrival duration using GPS data, *Transp. Res. Part E,* 56, 94-107

Ruan, M., Lin, J. and Kawamura K. (2012), Modeling urban commercial vehicle daily tour chaining. Trans. Res. Part E. 48, 1169-1184

Sharman, B. and Roorda, M. (2011), Analysis of freight global positioning system data: clustering approach for identifying trip destinations, *Transp. Res. Record,* 2246, 83-91

Sharman, B. and Roorda, M. (2010), Analysis tool to process passively-collected GPS data for commercial vehicle demand modelling applications. TRB–SHRP2 Presentation, Accessed Nov. 18, 2013, <http://onlinepubs.trb.org/onlinepubs/shrp2/C20/014AnalysisTool.pdf>

US Department of Transportation (2013), Accessed on Nov. 5, 2013, <http://www.rita.dot.gov/bts/sites/rita.dot.gov.bts/files/subject_areas/freight_transportation/html/freight_and_growth.html>

Zhao, W., McCormack, E., Dailey, D.J. and Scharnhorst, E. (2013), Using truck probe gps data to identify and rank roadway bottlenecks. *Journal of Transp. Eng.*, 139(1), 1-7

*The authors wish to thank Transport Canada and Shaw Tracking for providing the GPS dataset used in this paper*