EXPANDING TRUCK GPS-BASED PASSIVE ORIGIN-DESTINATION DATA IN IOWA AND TENNESSEE

Corresponding Author:

Vincent L. Bernardin, Jr., Ph.D.
Senior Consultant, RSG
2709 Washington Ave, Ste. 9
Evansville, IN 47714
Vince.Bernardin@RSGinc.com
Ph: 812-200-2351

Steven Trevino
Analyst, RSG
2709 Washington Ave, Ste. 9
Evansville, IN 47714
Steven.Trevino@RSGinc.com
Ph: 812-200-2351

Jeffery Short
American Transportation Research Institute
2060 Franklin Way, Suite 201
Marietta, GA  30067
JShort@trucking.org
Ph: 770-432-0628

Submitted: August 1, 2014

Word Count: 3,356 words + 8 tables/figures @ 250 words = 5,356 equivalent words
ABSTRACT
The availability of large samples of truck GPS data has presented a new, unprecedented source of information for understanding truck travel patterns and forecasting truck demand. This data is now being incorporated in the development of statewide models, beginning with Indiana’s in 2012 and now with Iowa and Tennessee’s. In order to use the data in modeling and forecasting, it is necessary to expand the GPS-based sample data to represent all truck movements. This presents a challenge, since the sample is not randomly drawn and cannot therefore be presumed to be representative. In particular, this paper presents confirmation that short-haul movements, while present in the data, are under-represented. Without correcting for this, it is not possible to produce accurate information regarding average trip lengths, etc., or to accurately forecast future truck activity. This paper presents on-going work to understand the representativeness of the American Transportation Research Institute’s (ATRI) truck GPS data and develop methodologies to produce factors for expanding it to ensure it is representative of truck travel patterns in general.

Keywords: ATRI, Truck GPS, Passive Data, Big Data, Expansion, Weighting
INTRODUCTION AND OBJECTIVES

The availability of large samples of truck GPS data has presented a new, unprecedented source of information for understanding truck travel patterns and forecasting truck demand. This data is now being incorporated in the development of statewide models, beginning with Indiana’s in 2012 and now with Iowa and Tennessee’s. However, in order to use the data in modeling and forecasting, it is necessary to weight or expand the GPS-based sample data to represent all truck movements.

The American Transportation Research Institute (ATRI) is a 501(c)3 not-for-profit research organization. As an independent part of the American Trucking Associations (ATA) Federation, it receives truck GPS position data from many of its members. The primary source of the data is on-board communications and navigation equipment installed on commercial trucks. The resulting dataset is extremely large and growing, currently nearing 1 billion data points per week, representing several hundred thousand individual trucks out of the total 2.4 million trucks registered in the US. (1) ATRI compiles the data as part of the Freight Performance Measures Initiative, an effort sponsored in part by the Federal Highway Administration.

ATRI’s truck data has been used for nearly a dozen years in many analyses to identify truck bottlenecks, congestion and speeds on major highway freight corridors across the country. More recently it was used to better understand truck origin-destination patterns as part of a feasibility study of dedicated truck lanes on an 800-mile section of I-70 through Missouri, Illinois, Indiana and Ohio with funding from FHWA’s Corridors of the Future program. The
GPS data proved to be extremely useful in understanding the patterns of truck movements utilizing the corridor as well as “ground-truthing” and improving travel demand forecasts. However, the data was not formally incorporated in any models.

The Indiana Statewide Travel Demand Model was the first travel forecasting model to formally incorporate the data when it was updated in 2011-2012. (2) The Indiana Department of Transportation obtained an eight week sample of ATRI’s truck GPS data drawn from each of the four quarters of 2010. The resulting sample contained over 16 million records representing over 2 million trips by over 300,000 trucks. The data was analyzed and while it was determined that the ATRI data clearly includes short-haul movements, it was determined that the preponderance of the data related to medium- and longer-haul truck trips and it was therefore used to improve the commodity-flow based truck movements in the model. The incorporation of the ATRI data into the model resulted in substantial improvements in the truck model validation, decreasing the root mean squared error (RMSE) against roughly 6,000 truck counts from 69.3% to 60.6% and decreasing the mean absolute percentage error (MAPE) from 74% to 42%. (3)

Despite the very positive results on the truck model performance, it was recognized that further improvement was likely possible if the representativeness of the ATRI data was better understood so that it could be expanded more realistically.

Inspired by the success in Indiana, both the Iowa and the Tennessee departments of transportation have acquired truck GPS data from ATRI in order to support the update of their statewide travel models. As part of these efforts, RSG has undertaken research with the support of ATRI, to study the representativeness of ATRI’s GPS truck data and develop more refined methods of weighting or expanding it for modeling purposes. In particular, the desire is to better understand both geographic and distance-based biases in the sample relative to total truck traffic.

**DATA AND METHODOLOGY**

Both Iowa and Tennessee obtained an eight week sample of trucks passing through their respective statewide model study areas drawn from each of the four quarters of 2012. For each model, a GIS file representing travel analysis zones (TAZ) was provided to ATRI. ATRI selected all truck GPS traces entering, exiting, traveling within or passing through the overall area for each of the sampled weeks. A data management and analysis software package was used to further prepare the dataset for integration into the truck trip table. Truck positions for each unique vehicle were sorted into a time series, and within each series each truck position was matched with the subsequent truck position to produce a set of truck position pairs. The geodetic distance between the first and second truck positions for each of the truck position pairs was then calculated. ATRI then replaces the precise GPS location data in its records with the TAZ. In addition to supporting the ultimate development of a trip table, this process also offers some benefit of further ensuring the anonymity of the data by associating truck positions with geographic areas far more generalized than a discrete latitude/longitude position – which could allow for the development of an address-specific customer list. The dataset is then reformatted so that each record represents the movement of a truck between GSP ‘pings’. ATRI then delivers a dataset containing an anonymous truck identifier, the distance between pings, the TAZ position of the beginning and ending ping and the timestamp of the beginning and ending ping (see Figure 2).
This dataset must then be processed and further reduced to represent trips between origin-destination pairs. This is done in two steps, first identifying for each pair of GPS pings or movement record, whether the truck was in motion or stopped. This determination is made based on complex criteria of a minimum travel speed and a minimum elapsed time and/or distance. The complex criteria are necessary to avoid including brief stops at traffic signals or brief repositioning movements within a single site (see Figure 3 for an illustration).

Once the moving records and stopped records are identified, the records are processed to identify the origin and destination for each sequence of moving records. When a stop record was found in the list, it signified the destination of the trip and the origin for the subsequent trip (see FIGURE 2). The result is a list of trips by origin-destination pair, which can be aggregated by origin-destination pair to produce a trip-table in flat/list format, which, in turn, can be read into a matrix format file by most travel modeling software.
The resulting trip table still must be cleaned for several reasons discussed in turn below. The largest issue was GPS positional errors, or “blips,” where the GPS location jumps from one location to another in a way that could not possibly represent a real movement (e.g., a change in position of 50 miles in a span of 30 seconds). Given the large size of the sample data (with over 3 million trips for Iowa), it was not necessary to invest large effort to correct these blips; rather, trips with such errors were simply identified and removed from the dataset. Moreover, a very conservative test of what qualified as a good trip or a very liberal identification of blips was used, again given the luxury of the abundant data. Although it may be ideal to attempt to correct and recover some of these trips with blips, simply removing them is safest course of action and was deemed reasonable given the very large sample size.

Some trips at the very beginning and end of each two week period were removed in order to avoid capturing trip fragments or partial trips in progress at the beginning or end of the period. If a truck is initially moving (no starting records of a stop) within the first hour of the global start time (Feb 01 @ 00:00:00) then those records were flagged. This time was determined to be a reasonable buffer by looking at the starting time distribution of all initial trips. Similarly, trucks that did not display a final stop and had movement within 3 hours of the global end time (Feb 15 @ 00:00:00) were flagged. After a look at ping length and ending trip distribution, it was assumed that if a truck displayed no pings for over 3 hours that it’s trip had ended. Less than 1% of the total trips were affected by either the stop time or end time filters.

FIGURE 4 One Week Sample of 1,000 Trucks' GPS Traces Passing through Memphis, TN
For each trip, its GPS calculated length was also compared to a centroid-to-centroid geodetic distance. Trips were flagged if the ratio was outside the bounds of 0.7 & 2.25. This was used to catch both blips that slipped through the initial filter as well as undetected stops and helped to confirm "clean" probable trips. This filter also resulted in the removal of less than 1% of the total trips.

A small number of trips appeared to start and end in the same zone (an intrazonal trip), but with unreasonable VMT. Looking into individual pings, it appears that these trips went through several zones and made a large circular trip. This seemed be from either a brief trip outside the model or in many cases an undetected stop. Intrazonal trips greater than 30 miles were generally removed.

The final effort is on the expansion of this resulting raw trip table. The simplest method is to scale the trip table to reflect the total number of truck trips or truck VMT. However, this fails to account for differences in the portion of the universe of trucks represented in the sample. For example, it is known that short-haul movements, while present in the data, are under-represented; without correcting for this, it is not possible to produce accurate information regarding average trip lengths, etc. There was no initial hypothesis as to whether there were any geographic biases in the data, but it was considered a possibility, since the sample is not randomly drawn. If there were geographic biases it would be important to correct for this to avoid distorting the spatial distribution of trips.

The approach taken in this effort begins by simply scaling the raw ATRI trip table to represent the proper amount of truck VMT. Then, origin-destination matrix estimation (ODME) algorithms are applied using truck counts on the network and the scaled trip table as a seed. The results of the ODME are not used directly, but analyzed to identify significant patterns. Once patterns of over- and/or under-representation have been identified, a categorization scheme was developed to address these systematic issues and average adjustment factors for each category were calculated from the ODME results judged most reliable. The final expansion factors were then the product of the scaling factor and the adjustment factors. While ODME algorithms are an imprecise method, they provide a consistent, structured and logical framework for relating the information contained in truck count data to origin-destination patterns. Truck counts, while also imperfect, are widely available and believed to be the best source of unbiased data on total truck traffic.
The iTRAM highway network includes 84,090 links in Iowa, representing 28,563 miles of roadway, of which 36% have a truck count or count derived estimate of truck traffic (links with truck counts shown in red in Figure 3). This provided a rich dataset for ODME analysis and expansion.

While ODME algorithms are an imprecise method, they provide a consistent, structured and logical framework for relating the information contained in truck count data to origin-destination patterns. Truck counts, while also imperfect, are widely available and believed to be the best source of unbiased data on total truck traffic.

**MAJOR RESULTS**

The Iowa dataset included over 50 million movement records by over 135,000 individual trucks making over 3.1 million trips, including over 60 million truck miles of travel observed in Iowa over the 56 days in the sample. This translated into 1,083,152 truck VMT in Iowa on a daily basis. Total daily truck miles of travel in Iowa are estimated by Iowa DOT to be 10,731,507 based on 2012 HPMS data. This represents an average, overall sampling rate of roughly 10.1%, corresponding to a scaling factor of 9.908. Since the sample was for 56 days, however, this means that the raw daily OD table was produced by factoring the total down by 0.1769 to represent a daily number.

The raw OD table was used as seed matrix for ODME and the results were analyzed to evaluate possible biases based on geographic regions or trip length. Although there are some anecdotes of regional differences in ATRI sample coverage around the US, no evidence of systematic geographic differences in sample coverage were identified in Iowa or its immediate
halo area. Figure 6 illustrates the OMDE implied weights within Iowa and clearly demonstrates there are no systematic biases. It is still possible that geographic or regional biases exist in other parts of the US, but in and around Iowa this does not appear to be an issue. Therefore, geography was not used in the weighting and expansion scheme.

The analysis did, however, confirm suspected bias towards longer haul truck trips over short haul truck trips. At the same time, the data demonstrated that there were still a significant number of short haul movements in the date, such that they could be reasonably weighted to produce a more representative dataset. Figure 7 compares truck trip length frequency distributions from the raw ATRI data and resulting from ODME. It can clearly be seen that trips less than 60 miles in length are factored up in the ODME process, while trips greater than 60 miles in length are factored down very slightly.
A weighting scheme was therefore developed based on trip length, and can be seen in Figure 8. Short haul truck trips (<60 miles) were weighted up while long haul trips (>60 miles) were slightly weighted down. The resulting weights ranged from slightly over 0.7 to just over 2.5; the latter value only applying to truck trips less than 10 miles in length. Trips even just 10–20 miles in length only had to be factored up by less than 1.8. While this does indicate the need to correct the data for this bias, it also suggests that a simple and reasonable weighting scheme can produce reasonably good results, explaining roughly half the variation in ODME weighting.

**FIGURE 7** Comparison of Truck Trip Length Frequency Distributions in Raw ATRI data and from ODME

**FIGURE 8** Weighting Scheme for Iowa ATRI Data based on Trip Length
The (scaled) raw, weighted and ODME ATRI trip tables were assigned to the iTRAM highway network for testing purposes only, to understand how well these trip tables correspond to the truck counts on iTRAM’s network. The final iTRAM truck assignments will incorporate a variety of different data sources including information derived from the commodity flow model as well as from special truck generator datasets developed with data from Leonard’s Guide, FleetSeek, NMFTA, etc. These test assignments, which are based only on the ATRI data, should therefore not be confused with final truck model results and only provide some initial indication of what those final results may be.

When assigned to the iTRAM highway network, the (scaled) raw ATRI data resulted in a root mean square error (RMSE) of 116% versus the 63,044 truck counts in the network. Several things should be noted with this. Truck assignment results invariably have higher error associated with them compared to total assignment results for several reasons, first including less accuracy in truck counts compared to total traffic counts partly owing to the greater significance of axle adjustment factors, etc., and generally the greater uncertainty of various count technologies for producing classification counts. Second, relative or percent errors are larger for trucks since total truck truck volumes (the denominator by which the statistics are normalized) are smaller; because of this, much larger percent errors can correspond to small absolute errors in terms of number of trucks. Third, assignment errors over large geographies, i.e., for statewide models, are higher than for urban scale models, both for total traffic as well as for trucks due to their coarser representation of the network. Fourth, in the case of the iTRAM network, it is clear that some portion of the truck “counts” are in fact estimates rather than actual counts, so there is greater error associated with these estimates. Thus, although the assignment error for the raw ATRI data is much higher than typical errors associated with total traffic assignments in urban models, it is not necessarily a bad result for a statewide truck assignment.

The ODME trip table from the ATRI seed resulted in an RMSE of 58%. This is comparable to the ODME based truck assignment results from the previous version of iTRAM which had an RMSE of 81%. This indicates that the ATRI data can ultimately help produce a substantially improved truck model compared to the previous model. However, final assignment results will be a function of other data, as well and may or may not incorporate ODME adjustments based on ATRI and/or other data.

The weighted ATRI trip table resulted in a RMSE of 92%. This means that the fairly simple distance-based weighting scheme was able to explain or accomplish more than 41% of the improvement from ODME adjustments. This is a very positive result, validating the approach and allowing us to understand and verify the reasonableness of an important part of the ODME adjustments. This result lends good support to Iowa’s ultimate model, whether it makes use of the weighted ATRI data directly or ODME adjustments informed by the ATRI data.

CONCLUSIONS

Given the importance of freight planning, it is critical to have accurate freight planning tools and models, but this is impossible without high quality data. The ATRI GPS truck data provides a rich and growing source of information on truck travel patterns. However, it is still only a sample, and like any sample care and an understanding of its representativeness is required in expanding it. Failure to correctly account for the under- or over-representation of certain categories of trucks could lead to faulty analyses and false conclusions.

This research presents an important step forward in our understanding of the representativeness of ATRI’s truck GPS data and how to weight or expand it for use in
forecasting models and other analyses. With proper understanding and expansion techniques, ATRI’s truck GPS data stands to support a new generation of truck models with substantially better accuracy than earlier models and methods based on very limited data. The research presented here is critical to ensure that the data is used in appropriate way.

REFERENCES

