Estimating External Travel Using Purchased Third-Party Data

Prepared by:

Harvey J. Miller, The Ohio State University
Morton E. O'Kelly, The Ohio State University
Young Jaegal, The Ohio State University
William Bachman, Westat
Leta Huntsinger, Parsons Brinckerhoff
Greg MacFarlane, Parsons Brinckerhoff

Prepared for:

The Ohio Department of Transportation,

Office of Statewide Planning & Research

State Job Number 134877

May 2018

Final Report

U.S. Department of Transportation
Federal Highway Administration

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
FHWA/OH-2017-35		

4. Title and Subtitle	5. Report Date
Estimating External Travel Using Purchased Third-Party Data	**May 2018**
	6. Performing Organization Code
7. Author(s)	8. Performing Organization Report No.
Harvey J. Miller **Morton E. O'Kelly** **Young Jaegal**	
9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
The Ohio State University **Center for Urban and Regional Analysis** **1176 Derby Hall, 154 N Oval Mall** **Columbus, OH 43210-1361**	
	11. Contract or Grant No.
	SJN 134877
12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Ohio Department of Transportation **1980 West Broad Street** **Columbus, Ohio 43223**	**Final Report**
	14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

Archived travel data (ATD) derived from various private sources has attractive characteristics that suggest it can be a suitable replacement for traditional sample-based external travel studies or other similar origin-destination (OD) studies. The hope is that this new source of information will reduce or eliminate several negative characteristics of traditional methods. Before this hope can be realized, however, the new solution must be validated; this is the main intent of this research project.

17. Keywords	18. Distribution Statement
Origin-destination, OD data, roadside survey, data accuracy, big data, purchased OD data, GPS trace data, LBS data	**No restrictions. This document is available to the public through the National Technical Information Service, Springfield, Virginia 22161**

19. Security Classification (of this report)	20. Security Classification (of this page)	21. No. of Pages	22. Price
Unclassified	**Unclassified**	**102**	

Estimating External Travel Using Purchased Third-Party Data

Prepared by:

Harvey J. Miller, The Ohio State University

Morton E. O'Kelly, The Ohio State University

Young Jaegal, The Ohio State University

William Bachman, Westat

Leta Huntsinger, Parsons Brinckerhoff

Greg MacFarlane, Parsons Brinckerhoff

May 2018

Prepared in cooperation with the Ohio Department of Transportation

and the U.S. Department of Transportation, Federal Highway Administration

The contents of this report reflect the views of the author(s) who is (are) responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the Ohio Department of Transportation or the Federal Highway Administration. This report does not constitute a standard, specification, or regulation.

Table of Contents

List of Tables

List of Figures

1 EXECUTIVE SUMMARY

1.1 Introduction

Transportation planning agencies have used cordon travel surveys empirically to determine the extent and character of personal and commercial travel entering, leaving, and passing through a defined study area. Traditionally, sample-based cordon surveys are conducted at all major study area access points and require stopping travelers and interviewing them about their trip origin, destination and purpose. This intercept approach requires extensive field staffing and is disruptive to travel. Other approaches are used that observe vehicle movement without stopping traffic (license-plate matching, Bluetooth sensing, and RFID sensing) but generally lack important trip characteristics needed to build predictive models. Neither approach meets the needs of cost effectiveness and value to transportation model development.

Technology advances over the last decade have progressed to the point where archived mobile phone and global positioning system (GPS) data can now be used to estimate trip characteristics for a significant percent of regional travel. Data based on mobile phone and GPS activity is being archived by private companies and mined to identify trips, typical trip ends, and other travel characteristics. The availability of aggregations of these data for purchase by transportation planning agencies offers a new alternative to the Ohio Department of Transportation (ODOT) and others for completing the mission typically addressed with traditional cordon travel surveys and similar OD studies. Before full acceptance, however, there is a need to study the potential bias, accuracy, applicability, and validation using systematic methods by a research organization. At the heart of this research is a timely question regarding the application of “Big Data” as a replacement for sample-based studies. This project is designed to answer these questions to the extent that ODOT can confidently conduct a new round of statewide cordon studies.

1.2 Objectives

Archived travel data (ATD) derived from various private sources has attractive characteristics that suggest it can be a suitable replacement for traditional sample-based external travel studies or other similar origin-destination (OD) studies. The hope is that this new source of information will reduce or eliminate several negative characteristics of traditional methods. Before this hope can be realized, however, the new solution must be validated; this is the main intent of this research project. This proposed project has three primary research questions that, when answered, will result in guidance for ODOT regarding future cordon travel survey methods, OD data collection methods, data purchases, data application methods, and model development strategies:

1. Does ATD offer any quality or sampling improvements or limitations that enhance or limit traditional travel demand forecasting model performance?
2. What ATD specifications should be applied to maximize value and minimize cost?
3. How can ATD be applied in traditional travel demand forecasting models to make maximum use of its strengths and minimize the impact of its limitations?

To answer these questions, data were acquired and analyzed from three ATD vendors. **Vendor A** data consists of generic vehicles (not distinguished, e.g., personal versus commercial vehicles). Vendor A derives their data entirely from mobile phone signal based on triangulation from towers during phone activity.   **Vendor B** data distinguishes between personal and commercial vehicles. This vendor derives data from navigation/traffic applications, extracting GPS tracks from users, as well as fleet/commercial vehicle GPS probe data.   **Vendor C** data comprises GPS trajectories from commercial vehicles.

1.3 Findings

1.3.1 Literature Review

The assessment of past related research findings provided direction regarding OD estimation methods and OD data management. These issues helped guide the research team with the initial analytical tests. Additionally, the research also revealed archived travel data issues that could be relevant to ODOT and the potential implementation of ATD. Among the most significant issues are:

1. Traffic Analysis Zones are likely too small for raw ATD. The resolution, accuracy, and value is better at larger aggregations. This issue can influence the cost of the purchased data. Higher resolution data may not provide any benefit and could cost substantially more. This does, however, necessitate a method to disaggregate ATD to the resolution of the transport analysis.
4. Due to the nature of identifying external trip ends and travel activity outside of a defined study area, the external catchment areas must be carefully considered.

1.3.2 Trip Length Analysis

The following findings are of significance to ODOT:

1. All ATD products matched well when compared to the 2011 model data for EE trip lengths. Vendor A and Vendor C data required significant cleaning to reach this conclusion due to the external catchment definitions used in developing the Vendor A product and the privacy protection techniques used in the Vendor C data.
2. Overall, the trip length comparison for IE/EI trips did not perform as well as EE trips.
3. Vendor A data underestimates short IE/EI trips (less than 5 minutes)
4. Vendor B data provided the best IE/EI comparison for personal vehicles
5. Vendor C data provided the best IE/EI comparison for trucks

1.3.3 Trip Purpose Analysis

The following findings are of significance to ODOT:

1. Only Vendor A provided estimates of trip purpose and person type that could be used in comparison to the 2008 survey
2. The Vendor A trip purpose estimates did not compare well with 2008 survey

1.3.4 External – External (EE) Flow Analysis

The following findings are of significance to ODOT:

1. All ATD products matched well when compared to the 2011 model data for EE trip flows when using the cleaned version of each product (see section 4.1.2).
2. Vendor B demonstrated the best EE flow comparison for personal trips
3. Vendor B demonstrated the best EE flow for commercial trips

1.3.5 External-Internal Trips Analysis

The following findings are of significance to ODOT:

1. None of the ATD sources compared well to the 2011 model data for EI trips.

1.3.6 Assessment of highway network traffic flows

The following findings are of significance to ODOT:

1. All of the ATD sources compared well to the 2011 model data
2. Vendor B performed the best for personal travel
3. Vendor B and Vendor C both performed well for commercial travel

1.3.7 Modeling impacts

The following findings are of significance to ODOT:

1. ATD from Vendor A and Vendor B produced similar results in a network assignment step, *after* transforming the ATD data via iterative proportional fitting to the flows at the external stations.

1.4 Recommendations

The primary objective of this research was to determine if ATD can be used to replace the costly and burdensome traditional cordon survey process in Ohio. Results from this analysis of Lima, Ohio paint a complicated picture where a confident and single recommendation is not possible. The analysis procedures indicated that Vendor B performed the best of the three products. Further, Vendor B has both a personal and commercial product that eases the integration effort. Arguments that prevent a full and confident recommendation of Vendor B (or any ATD product) is that the overall performance was not as strong as hoped and the analysis was limited to a single study area. Additionally, the ODOT Cordon Survey was conducted in 2007, the date of which may somewhat limit the results of direct comparisons. The original study questions and secondary study questions are as follows:

Does ATD offer any quality or sampling improvements or limitations that enhance or limit traditional travel demand forecasting model performance?

Based on the analysis of Lima, OH, the ATD products do not offer unambiguous quality improvements when compared with ODOT data. Sampling improvements with ATD are evidentially better since they do not share the spatial and temporal limitations of traditional surveys: EE, EI, and IE trip information can be gathered for an extended period of time, a clear advantage over traditional methods. However, there does appear to be bias in the ATD that is difficult to identify given the study parameters. The primary reasons for concern are noted throughout the research discussion but include poor results in comparing the survey station to zone flows, lack of trip purpose details, and product-specific issues. On the positive side, an understanding of the ATD sources and methods allows for the effective use of pre-processing techniques to improve performance. Section 5.2 provide specific details on these methods.

Vendor B performed better than the other solutions potentially due to the fact that their raw data has a better spatial resolution and it is therefore better at defining the exact external station entry and exit points.

What archived cellular data specifications should be applied to maximize value and minimize cost?

The specifications for each product may change based on offerings from each provider, but the following items should be considered when negotiating the purchase of a product:

1.4.1 Vendor A

Vendor A offers a wide variety of specifications. For the specific implementation of Vendor A data as a replacement of traditional cordon surveys and based on the Lima OH analysis, the following specifications are recommended to maximize value:

1. Do not opt for trip type variation – results of trip type comparison were inconclusive, however, comparisons between the survey data and Vendor A’s data showed significant differences. Subsequent analysis of large employment centers showed low numbers of work related trips by Vendor A. It is possible that Vendor A could improve their capabilities in a location with higher populations, higher market penetration, and improved capability to identify EE, IE and EI trip end locations outside of the study area. The current limited catchment area approach may be limiting their ability to estimate trips and trip types.

2. Do not opt for person type variation – results could not be verified, and the person type provided limited value to the traditional external travel model development.
3. Aggregate TAZs to larger districts – the TAZ level analysis showed significant variation from the survey. Reducing the number of zones reduces the cost and potentially improves the quality of the data. This also requires that trip disaggregation techniques be applied after the data purchase to assign district level trip information to specific TAZs.
4. Clearly define the catchment areas outside of the study area – large catchment areas along travel corridors will likely provide better data. Due to the nature of Vendor A's methods, it is expected that this is critical to capturing the external travel.

1.4.2 Vendor B

Vendor B offered fewer specification details and appeared to limit options based on their internal structure of their data. The pricing of the product also appeared more fluid, making it more difficult to determine the value of various specifications.

1. Aggregate TAZs to larger districts – the TAZ level analysis showed significant variation from the survey. Reducing the number of zones potentially reduces the cost and potentially improves the quality of the data. This also requires that trip disaggregation techniques be applied after the data purchase to assign district level trip information to specific TAZs.
2. Expand the temporal coverage of the data – This analysis relied on a single month of data, but there are potential, yet unproven, advantages to selecting multiple months or seasonal data collection periods.
3. Provide specific external station location details – Vendor B has good geospatial information that can be used to match to specific external stations. However, lightly traveled stations may show more bias as the expected raw data source penetration is limited.
4. Identify specific truck weigh stations and truck stops – Analysis of Vendor B commercial data in Lima suggested that their trip end identification techniques mis-identified weigh stations and truck stops as an origin or destination. While vehicles did actually stop at those locations and this is significant for modeling, a full understanding of where a truck trip was going was limited.

1.4.3 Vendor C

Vendor C offered fewer specification details and significant privacy protection restrictions on their data. Their data does not come fully processed into origin-destination details but is simply a trace of travel details using a time-limited window. Spatially, the travel trace is provided as a sequence of zone-IDs along with vehicle speed and timestamp. Vendor C was also the most inexpensive source due to its raw form and the nature of the organization providing the data. The following specifications are suggested if ordering Vendor C data:

1. Provide TAZs – Vendor C will use zones specific to your study area. As a default they use US Census block groups.
2. Provide external catchment areas as TAZs – Since Vendor C uses US Census block groups as default zone ID for identifying the location of trip details, there is a challenge in determining actual travel roads and entry/exit points of the study area. Providing a catchment area extending out from each external station eliminated this problem.

How can archived travel data be applied in traditional travel demand forecasting models to make maximum use of its strengths and minimize the impact of its limitations?

It is important to understand that all data sources have limitations and contain margins of error. One can consider ATD and surveys as two ends of a spectrum, where travel surveys can provide maximum detail on individual trips and ATD can provide a much deeper sample of trips from a global perspective.

An external trip model that fully utilizes the strengths of ATD may look different from current models based on intercept survey data. Because ATD providers' purpose imputation can be weak and underdeveloped (a finding of this and other studies), modeling internal/external trip attractions by purpose may not be possible or desired. Similarly, given the finding that larger districts and catchment areas result in more coherent results, it is prudent to eliminate external stations with very small flows (less than approximately 1k AADT), particularly if there are many such roads on the same side of the model.

If it is essential that an internal/external trip model capture trip purpose, then ATD alone may not be sufficient. On the other hand, current trip attraction models estimated on survey data typically have large standard errors and low predictive power because the variables available to predict trips (floor space, jobs) are only weakly correlated with trip making.

2 BACKGROUND / LITERATURE REVIEW

2.1 Quality measures for Origin-Destination trip table estimation from Archived Travel Data (ATD)

Mobile network operators collect locational information for many reasons, such as billing, troubleshooting, and continuous coverage of service (Caceres, Wideberg and Benitez*,* 2007; Yin et al.*,* 2015). This *archived travel data* (ATD) is a promising source for transportation studies. ATD has four advantages over traditional travel data collection methods such as household and roadside surveys: i) larger size of the data sample; ii) broader spatiotemporal coverage; iii) lower collection cost, and; iv) shorter update intervals (Caceres, Wideberg and Benitez*,* 2007; Calabrese et al.*,* 2013). Due to these merits, a growing number of transportation organizations are using ATD across a wide range of transportation studies (Iqbal et al*.*, 2014).

Recent studies also have drawn attention to the potential of ATD to the estimation of origin-destination (OD) trip tables. OD trip tables are an integral part of travel demand modeling and transportation planning but has relied heavily on the traditional, expensive data collection methodologies. Network operators have locational information with frequent updating in order to provide continuous service coverage. This locational information provides the basis for estimating OD trip flows for a study area using ATD (Caceres, Wideberg and Benitez, 2007).

Despite the potential of ATD for estimating OD tables, these estimates need to be validated before accepting ATD as an alternative to traditional data collection tools. In this respect, two critical issues are: i) the accuracy of OD information from ATD, and; ii) privacy protection. Although OD matrix estimation from ATD is a relatively new research topic, there is a growing literature on measuring the reliability of OD trip table from ATD (Hard et al*.,* 2014; Liu *et al.* 2014). Also emerging are statistical measures for comparing OD trip tables from different sources (Chen et al*.,* 2015; Gan, Yang and Wong, 2005; Yang, Iida and Sasaki, 1991). Several scholars are also addressing the privacy concerns that may be raised by using ATD in transportation studies (Chow and Mokbal, 2011; Lu and Liu, 2012; Ratti et al*.,* 2006).

2.2 O-D estimation from Archived Travel Data

2.2.1 ATD as source for OD estimation

Commercial companies collect location information of their customers or users for a variety of practical reasons such as billing and fast reaction to technical problems (Yin et al., 2015). Companies track and archive location information in order to guarantee that its users are continuously connected to the network in any area (Caceres, Wideberg and Benitez, 2007). Wireless sensors in mobile devices include GPS (Global positioning system), WiFi (wireless fidelity), Bluetooth and radiolocation capabilities. After being processed to make it difficult to re-identify users, and packaged by private data firms, ATD becomes available for transportation planners and researchers via purchase.

ATD has four major advantages over traditional travel survey data sources (Caceres, Wideberg and Benitez, 2007; Calabrese et al., 2013). First, the sample size of ATD is typically much larger than available via traditional travel survey methods via as mailings, phone calls or intercept surveys. Second, ATD can also be collected over a broader range of locations and times than traditional survey data, allowing more blanket coverage of the travel pattern in a study area. Third, the cost of data collection is lower than other methods because it makes use of pre-installed infrastructure. Fourth, ATD can be updated in shorter time interval compared to traditional ones. Accordingly, a growing number of transportation studies are paying attention to the potential of ATD to address a range of research topics including modeling, visualization, and pattern recognition of human mobility (Iqbal et al., 2014).

ATD can serve as input for estimating OD flows for travel demand modeling and transportation planning models. The traditional data collection tool for OD estimation for external trips is the cordon survey. A sample-based cordon survey interviews travelers passing through a study area about trip characteristics such as origin, destination and trip purpose, and so on. While this type of survey produces a sufficient level of trip information needed to build travel demand models, it is expensive due to the high cost of field staffing. Additionally, cordon surveys hinder the flow of traffic because travelers are required to make roadside stops to reply to the survey. As technologies advance, new cordon methods have emerged that do not require stopping traffic; these include automated license plate recognition and traffic sensing via wireless technologies such as Bluetooth and RFID. However, these types of data collection tools provide little information on trip characteristics that are central to model building. As noted above, ATD has a set of advantage over these methods regarding cost-effectiveness and adequacy.

2.2.2 OD demand estimation from ATD

The most frequently used localization methods for collecting ATD is via the communication network. This is typically a network of base transceiver stations (BTS) distributed over a given region to provide the best possible radio coverage (Smoreda, Olteanu-Raimond, and Couronné, 2013: 747). The basic positioning methodology is as follows (Caceres, Wideberg and Benitez, 2007; Zhang et al., 2015). A communication network service area consists of a set of smaller hexagonal regions comprising BTS service areas. The group of adjacent regions to a station forms a 'location area' (LA): this can comprise as many as several hundred individual service areas. Four types of 'signaling events' trigger the localization procedure: i) communication events (call, SMS and internet service); ii) handover (i.e. service region changes by the movement during communication events); iii) location area update (LAU, i.e. tracking inactive devices when they move across LA border); iv) periodic location update (PLU). The former two events occur when the device is communicating while the latter two events occur irrespective of whether the device is in use or not. Throughout this process, the device's location data are recorded automatically and immediately to operator's database.

Since the network-based positioning method approximates a device's location based on the locations of the BTS[1], localization accuracy is poorer than GPS and WiFi-based methods. However, they have two definite advantages over other methods. First, the method can be implemented on older devices not equipped with GPS and WiFi capabilities, allowing more devices to be used as personal traffic sensors. More importantly, the infrastructure needed for the method is already installed by network operators. The remainder of this section focuses on ATD collected by network-based positioning methodology.

2.2.3 Methods for deriving OD demand information from ATD

ATD from network-based methods can be divided into two categories: i) mobile probe data (also called technical network logs); and, ii) call detail records (CDR). Mobile switching centers (MSCs) collect data via mobile network probes for technical management of the communication network. An MSC is a middle layer of network management controlling the network switching subsystem and recording the tracking information. Each LA has its own MSC, and all tracking information is stored in the database in MSC. Mobile probe data usually contain timestamp, service region ID and location area code of all four kinds of signaling events listed above (Smoreda, Olteanu-Raimond, and Couronné, 2013).

Recent studies of mobile probe data for OD demand estimation include Caceres, Wideberg and Benitez (2007), Zhang et al. (2015) and Larijani et al. (2014). Caceres, Wideberg and Benitez (2007) used a simulator tool emulating a real-world communication network and vehicles with mobile devices onboard. In their simulation setting, the network simulator tracks the location of simulated vehicles in the same manner as a real MSC. After dividing a day into specific time intervals, they retrieved individuals' trajectories by assuming that the first register of a vehicle during the time window is the origin location and the last one is the destination location. Similarly, Zhang et al. (2015) conduct simulation experiments to propose an estimation method for daily OD demand from mobile probe data. In their study, the location of the first signal event in the morning is identified as the trip origin. Then, the trip destination is determined as the location with the longest duration and distance from the origin. Unlike to these two studies, Larijani et al. (2014) estimate OD demand from actual mobile probe data of 1.4 million phone users in Paris, France. They also applied a temporal window for extracting the daily trajectories.

CDR is a method for billing where a mobile operator records each client's history of device usage. Similar to mobile probe data, CDR usually contains both timestamps of each event and the spatial location of the service region in which the customers connect to the network (see Iqbal et al. (2014: 66) for an example template of CDR data). However, unlike mobile probe data, CDR is only associated with communication event and handover which occur when cell-phone is in use.

Two recent studies that examine CDR as a source for OD trip tables are Iqbal et al. (2014) and Alexander et al.(2015). Iqbal *et al.* (2014) proposed a method for OD demand estimation by using CDR data of 6.9 million users in Dhaka city in Bangladesh. In order to extract individual trajectories from CDR, they assume that two consecutive locations within a specific time window are nodes on the user's trip. The time window they used was between 10 minutes and 1 hour. Similarly, Alexander et al. (2015) extract a group of recorded location of each individual that are spatially and temporally clustered. These clusters, called 'stays,' become 'candidate' nodes on each individual's trip. Based on these stays, they derived trajectory information of each user's daily trips from the CDR of 2 million users in the Boston metropolitan area. One thing to note is that they attempted to take uncertainty into account. Instead of taking the observed departure time for granted, they randomly generated the departure time of a trip by using the trip distribution of residents in consolidated metropolitan statistical areas (CMSA) obtained from National Household Travel Survey (NHTS).

[1] Some mobile operators also provide estimates of location based on triangulation algorithms (see Alexander et al. 2015).

Both mobile probe and CDR data have advantages and disadvantages (Smoreda, Olteanu-Raimond, and Couronné, 2015). Mobile probe data are an adequate source of individual mobility data because the dataset has location information regardless of whether the device is in use. However, since mobile probe data are collected and managed for technical reasons, the data can be nonstandard and unwieldy for other purposes such as travel demand analysis. Also, mobile probe data are specific to each MSC and there is no native function to merge them across MSCs. On the other hand, CDR is more widely used in transportation research because of its larger sample size and standard format. However, it contains less information on mobility patterns. For example, Iqbal et al. (2014) pointed out that consecutive CDR locations often have long time gaps.

2.2.4 Adjustment factors for initial O-D matrix

OD demand data obtained directly from ATD contains only a sample of clients from a single network operator. Thus, initial OD demand information needs to be scaled up to represent all vehicle traffic or population in the study area. A variety of adjustment methods has been used to achieve this goal. Caceres, Wideberg and Benitez (2007) develop an adjustment factor called device per vehicle equivalent (DVE) to convert the flow of mobile phone to those of traffic. Three parameters are considered in the calculation of DVE: the number of occupants in the vehicle, the market share of the network operator, and the likelihood that a device is switched on.

Zhang et al. (2015) utilize an adjustment measure called vehicle-per-device (VDE) equivalent factor. VDE is more sophisticated than DVE in two ways. First of all, VDE considers that the proportion of phone users correlates with socio-economic characteristics of the population in the study area. Zhang et al. (2015) use simulation to estimate the conditional probability of mobile phone ownership using income and age distribution obtained from census, instead of using mere market penetration and market share. A second advantage of VDE is an adjustment factor considering the posterior information of trip trajectories crossing LA boundaries and therefore being detected. Specifically, they classified the probe trajectories into three groups according to the number of LAs through which it passes: at least two, just one and zero, and develop adjustment factors for each group.

Unlike the two studies aimed at traffic demand, Alexander et al. (2015) scaled up the initial OD information to estimate the population in a study area. In their method, a 'home' location is identified as the stay visited the most on weekends and weekdays between 7 pm and 8 am. Then, an expansion factor is calculated as the ratio of the census population to the number of home locations.

As discussed below, there was no basis to adjust the acquired ATD to generate population estimates in this study. Therefore, the analysis is based directly on the ATD. This raised some issues regarding differences in sample sizes among the datasets.

2.3 Quality measures for OD matrix from ATD

2.3.1 Validating OD matrices derived from ATD

In the previous section, a variety of recently developed methods for OD estimation from ATD were reviewed. These methods consist of two steps: i) constructing individual trajectories from the devices' recorded locations; ii) applying adjustment factors to the initial OD information for scaling up to traffic demand patterns in the study area. However, despite the merits of ATD and adjustment methods, ATD also has several drawbacks as a source for OD demand estimation (Calabrese et al., 2013). First, ATD is not based on a random sampling frame. A typical ATD dataset comprises a choice-based sample of clients using the service of a single network operator. The adjustment factors discussed previously allow scaling these data to population levels, but this may not resolve possible biases in the choice-based sampling. Second, ATD is not designed for travel modeling purposes. Operators collect ATD mainly for technical management and billing, making it difficult for researchers not only to deal with some ATD but also to extract accurate OD information because it lacks important information needed for this estimation. Third, there is the intrinsic inaccuracy of location information in ATD because recorded locations only resolve to the BTS location to which a device connects. For example, Kwan (2016) demonstrates that the trajectories of mobile phone users can be estimated differently according to the different spatial configurations of cell towers in the study area.

All these three aspects of ATD suggest that OD information from ATD is subject to error. Therefore, the accuracy of OD information from ATD needs to be quantified and validated before being used as a substitute for traditional data collection.

2.3.2 Statistical error measures for OD matrix

2.3.2.1 Absolute and relative error measures

When a true or "ground-truth" OD table is available, the accuracy of an estimated OD trip table can be measured by using the differences. Chen et al. (2015) and Gan, Yang and Wong (2005) discuss four statistical measures used for absolute and relative error in the literature:

Mean absolute error

$$\mathrm{MAE} = \frac{1}{|RS|} \sum_{rs \in RS} |q_{rs} - \bar{q}_{rs}| \tag{1}$$

Root mean square error

$$\mathrm{RMSE} = \sqrt{\frac{1}{|RS|} \sum_{rs \in RS} (q_{rs} - \bar{q}_{rs})^2} \,/\, \frac{1}{|RS|} \sum_{rs \in RS} q_{rs} \tag{2}$$

Relative error

$$\mathrm{RE} = \sqrt{\frac{1}{|RS|} \sum_{rs \in RS} (\frac{q_{rs} - \bar{q}_{rs}}{\bar{q}_{rs}})^2} \times 100 \tag{3}$$

Total demand deviation

$$\mathrm{TDD} = \frac{\sum_{rs \in RS} q_{rs} - \sum_{rs \in RS} \bar{q}_{rs}}{\sum_{rs \in RS} \bar{q}_{rs}} \times 100 \tag{4}$$

where q_{rs} and $\bar{q}_{rs}$ are estimated and ground-truth flow between origin *r* and destination *s* respectively; RS is set of all OD pairs; $|RS|$ is the number of all OD pairs in the network. These statistical measures compare each element in two matrices and average the error or deviation of estimates from the true traffic volume.

Another useful measure of relative error is the coefficient of determination or R-square measure.

The Geoffrey E Harvers method (GEH) is another measure widely used in the United Kingdom for measuring the quality of traffic demand model estimate for each OD pair (Transport for London, 2010):

$$\mathrm{GEH} = \sqrt{\frac{(q_{rs}-\bar{q}_{rs})^2}{0.5(q_{rs}+\bar{q}_{rs})}} \quad (5)$$

GEH is the product of the square root of the absolute error, $\bar{q}_{rs} - q_{rs}$, and the relative error, $(\bar{q}_{rs} - q_{rs})/0.5(\bar{q}_{rs} + q_{rs})$. One property of GEH is that it puts more weight on larger flows than smaller ones. Chitturi et al.(2014) pointed out that, even if two OD pairs shows the same level of percentage difference, the one having larger absolute error is more important in the whole traffic system. For this reason, Chitturi et al.(2014) adopted GEH formula for comparing O-D information from Bluetooth signal data and a ground truth OD table. Blogg et al. (2011) is also one example of studies using GEH statistics to measure the quality of OD trip table from Bluetooth sensing technology.

Table 1 provides threshold values used for interpretation of GEH values, with numeric examples based on flow size between an OD pair (Van Vliet, 2013). According to the rule of thumbs in Table 1, the O-D pairs with a GEH value less than five would be evaluated as an acceptable fit. Additionally, Table 1 shows that GEH value is more elastic to change of the percentage error of the larger flow (i.e. 4000) than that of the smaller one (i.e. 500), demonstrating that GEH puts more emphasis on larger flows. The British Highways Agency Design Manual for Roads & Bridges (DMRB) requires the GEH to be less than 5 for at least 85% of the OD pairs (Chitturi et al. 2014).

Table 1 GEH interpretation and examples

GEH values	Rating	Numeric examples	
		4000	500
1.0	Excellent	+/- 65 (1.6%)	+/- 25 (5.0%)
2.0	Good	+/- 130 (3.3%)	+/- 45 (9.0%)
5.0	Acceptable	+/- 325 (8.1%)	+/- 120 (24.0%)
10.0	Unacceptable	+/- 650 (16.3%)	+/- 250 (50.0%)

GEH is an intuitive, empirical measure that is not grounded in statistical theory (Van Vliet, 2013). If the denominator of GEH is changed to either $\bar{q}_{rs}$ or q_{rs} instead of $0.5(\bar{q}_{rs} + q_{rs})$, the GEH^2 will have the same value as a chi-square statistics. However, the chi-square statistic will be highly likely to indicate that the two OD matrices are significantly different. Nevertheless, Van Vliet (2013) finds GEH useful for transport modelers because their focus is primarily on the applicability of the estimated OD trip table even if fit is not statistically rigorous.

2.3.2.2 Estimated error measures

In many applications, the true OD trip table is unknown. In this case it is misleading to use the absolute and relative error measures discussed above. Instead, estimated error measures for OD should be used. The error measures discussed in this section were designed to measure the error of estimated OD matrices from observed traffic flow within links in the network. They are based on the assumption that the link flows and link-use proportions (i.e. the proportion of trip demand of an OD pair using a particular link) are error-free. Therefore, the methods presented in this section are not directly applicable to OD trip matrix constructed based primarily on trip trajectories estimated from ATD. However, ATD can also be used to estimate link flows or enhance estimates based on traditional link counts (Caceres et al. 2012). Thus, the statistical measures in this section can be used for measuring errors between OD tables based on observed link counts and OD tables based on ATD-based link flow estimates.

Maximum possible relative error (MPRE). The maximum possible relative error (MPRE) estimates the upper bound of the relative error for an estimated OD matrix (Yang et al., 1991). MPRE assumes that: i) observed traffic link count, $\bar{v}_a$ (i.e. the link count on a link, a; $a \in \bar{A}$), is error free, and; ii) the route choice proportion, P_a^{rs} (i.e. the proportion of demand between of an origin, *r*, and a destination, *s* using link *a*) is accurately specified. Therefore, both the estimated OD trip table and the true table must reproduce traffic counts when assigned based on the link usage proportion:

$$\sum_{rs \in RS} P_a^{rs}\, q_{rs} = \overline{v_a}\ ,\ \forall\, a \in \bar{A}, \tag{6}$$

$$\sum_{rs \in RS} P_a^{rs}\, \overline{q_{rs}} = \overline{v_a}\ ,\ \forall\, a \in \bar{A}, \tag{7}$$

Let $\lambda_{rs} = (\overline{q_{rs}} - q_{rs})/q_{rs}$ denote the relative error between two tables. The OD trip demands, $\overline{q_{rs}}^*$ from which the estimated demands deviate the most can be obtained by solving the following quadratic program (Chen et al. 2015).

$$\text{Maximize } \sum_{rs \in RS} \lambda_{rs}^2 \tag{8a}$$

subject to

$$\sum_{rs \in RS} \lambda_{rs} P_a^{rs} q_{rs} = 0\ ,\ \forall a \in \bar{A} \tag{8b}$$

$$\lambda_{rs} \geq -1,\ \forall a \in \bar{A} \tag{8b}$$

then, the MPRE is given by

$$\text{MPRE} = \sqrt{\frac{1}{|RS|} \sum_{rs \in RS} \left(\frac{\overline{q_{rs}}^* - q_{rs}}{q_{rs}}\right)^2} \tag{9}$$

Expected relative error (ERE). MPRE considers only the worst case of a true OD trip table. The solution to the quadratic program in equation (8a) is a vertex within the feasible solution region formed by the linear constraints in equation (8b) and (8c). However, a true OD trip matrix can occur at any point in the feasible region. To address this issue, Gan, Yang and Wong (2005) proposed a new estimation error measure called the expected relative error (ERE). Their approach uses random sampling within the feasible solution region. Once a sufficient number of samples is obtained by random sampling, the error is computed based on both the probability of each sample and the deviation of the estimated table from the samples.

Any feasible OD trip tables, ***q***, of a certain OD estimation problem can be expressed as a convex combination of the extreme points (or vertices), $k \in K$, of feasible region.

$$\boldsymbol{q} = (q_{rs})_{rs \in RS} \text{ with } q_{rs} = \sum_{k \in K} \propto^k \, q_{rs}^k \tag{10}$$

$$0 \leq \propto^k \leq 1 \text{ and } \sum_{k \in K} \propto^k = 1.0 \tag{11}$$

A feasible OD matrix is sampled by randomly generating a set of $\propto = (\propto^k)_{k \in K}$ satisfying the constraint in equation (11). Let $\boldsymbol{q}^t, t \in T$ denote the sample OD matrix correspond to sample *t.* Then, the ERE is calculated as the sum of all product of likelihood of a sample t, $\Pr(\boldsymbol{q}^t)$, and the relative deviation of the estimated OD table from the sample matrix, Rel_t (Chen et al.*,* 2015; Gan et al.*,* 2005).

$$\text{ERE} = \sum_{t=1}^{T} \Pr(\boldsymbol{q}^t) \cdot Rel_t \tag{12}$$

$$Rel_t = \sqrt{\frac{1}{|RS|} \sum_{rs \in RS} (\frac{q_{rs}^t - q_{rs}}{\overline{q_{rs}}^*})^2} \tag{13}$$

the probability of the occurrence of a sample matrix $\boldsymbol{q}^t$ is given by (Chen et al.*,* 2015):

$$\Pr(\boldsymbol{q}^t) = \frac{(\sum_{rs} q_{rs}^t)!}{\prod_{rs} q_{rs}^t!} \cdot \prod_{rs} (\frac{\hat{q}_{rs}}{\sum_{rs} \hat{q}_{rs}})^{q_{rs}^t} \tag{14}$$

where $\hat{q}_{rs}$ is the trip demand obtained from a reference matrix, which is an existing OD trip table usually obtained from a former study or road side survey.

Total demand scale (TDS). Bierlaire (2002) proposes a quality measure for OD trip tables called total demand scale (TDS). Basically, the TDS is the gap between the maximum value (φ_{min}) and the minimum value (φ_{max}) of the total trip demands for feasible OD matrices.

$$\text{TDS} = \varphi_{max} - \varphi_{min} \tag{15}$$

One can obtain the two extreme values by solving the following linear programs (Chen et al.*,* 2015; Bierlaire, 2002).

$$\varphi_{max} = \max_{\boldsymbol{q}} \sum_{rs\in RS} q_{rs} \text{ and } \varphi_{min} = \min_{\boldsymbol{q}} \sum_{rs\in RS} q_{rs} \tag{16a}$$

subject to

$$\sum_{rs\in RS} P_a^{rs} q_{rs} = \bar{v}_a \text{ , } \forall a \in \bar{A} \tag{16b}$$

$$q_{rs} \geq 0, \ \forall rs \in RS \tag{16c}$$

Bierlaire (2002) suggested that the value of TDS can be interpreted in three ways. First, if TDS equals zero, the total demand of the estimated OD trip table is correct, meaning that the estimation error is attributable solely to the way how the total demand is assigned to each OD pair. Second, if TDS is greater than zero, TDS becomes a measure of the range of the total trip demand of the feasible OD tables. In this case, both the total demand and the repartition of demand can be a cause of the estimation error. Thirdly, if TDS is positive infinite, it indicated that the link flow data used in the estimation violate the OD covering rule, meaning that the flow of some OD pairs is not captured at all by any of link counts. One thing to note is that, similar to TDS, MPRE is also infinite if the link counts data do not satisfy the OD covering rule.

2.3.3 Statistical measures for comparing trip profiles

Both actual and estimation error measures presented in the previous two sections consider the demand of each OD pair for quantifying the quality of the OD trip table directly. Another approach to this issue is comparing the profile of trip characteristics obtained from the estimated and the ground truth table. Trip characteristics can include trip purpose, trip length, the timing of departure and arrival, and trip type (e.g. external to internal, internal to external and external to external).

Correlation statistics are frequently used in the literature for providing an indication of the goodness of fit of the profile from the estimated OD trip table. For example, Liu et al. (2014) conducted a correlation analysis to demonstrate that the estimated profile of trip pattern (e.g. home to work or school and home to non-mandatory activities such as shopping, social visit or sports) from mobile phone data are consistent with the actual profile obtained from a survey.

Another viable option is the Kolmogorov–Smirnov test (K-S test). K-S test statistic is “a means of testing whether a set of observations are from some completely specified continuous distribution” (Lilliefors, 1967). It is also used to test whether two samples are generated from the same distribution. A recent example of the use of K-S test is the study of Hard et al. (2015). In the research, they compare trip length frequency estimated from different data collection tools including Bluetooth, mobile phone, and external survey to check whether the newer data are adequate as a substitute for the traditional data.

2.4 Privacy protection strategies for ATD

2.4.1 Locational privacy

Another important issue related to the use of ATD is privacy protection. Network operators and data firms related to producing and selling of ATD are using a variety of procedures for ensuring that their customers’ information is never re-identified by researchers. Privacy protection strategies, however, present a challenge to ATD providers because it may cause a loss of information that lowers the quality of ATD as a source for analytic purposes (Yin et al., 2015). In this respect, Lu and Liu (2012) warn that privacy protection can be compromised if data sellers put more emphasis on providing detailed information for analytic accuracy. This tradeoff between privacy protection and data utility implies that both data provider and consumer should be aware of the potential re-identification risk and the characteristics of privacy protection procedures applied for the ATD. For data providers, the challenge is to ensure that individuals cannot be re-identified while producing an analytically valuable data set. For researchers, the challenge is to take into account the intrinsic errors that may be caused by the privacy protection methods.

Discussions on privacy concerns of location data have been intensified along with the advance of pervasive location awareness technologies. Chow and Mokbel (2011) pointed out that the type of location information that continuously updated to a service provider such as ATD raises more serious privacy concerns than the snapshot location information. This is because the individual's trajectory can be inferred by investigating spatiotemporal dimension of a set of one's location information. They classified privacy concerns related to massive location data into three categories: i) data privacy; ii) location privacy, and; iii) trajectory privacy. Data privacy is related to the possibility of identifying each individual from de-identified microdata. Usually, a published microdata element is de-identified by removing unique identifiers such as name and unique registration numbers (e.g. social security numbers). However, one can re-identify individuals without a unique identifier by considering the combination of non-identity variables such as zip code, gender, and date of birth. These non-identity attributes are called quasi-identifiers. Location and trajectory privacy are concerned with the possibility of using inferred location and trajectory information as a quasi-identifier respectively. In the following section, two techniques widely accepted as effective for privacy protection of location data are introduced.

2.4.2 Privacy protection procedures for location data

2.4.2.1 k-anonymity

The *k*-anonymity principle is the most widely adopted technique for privacy protection in data science. Sweeny (2002) proposed k-anonymity protection model as a formal framework for preventing disclosures of sensitive personal information from publicly released data. A released dataset satisfies the *k*-anonymity requirement "if the information for each person contained in the release cannot be distinguished from at least *k*-1 individuals whose information also appears in the release" (Sweeny, 2002). In other words, the principle constraints that each unique combination of values comprising quasi-identifier must occur at least *k* times in the dataset. Specifically, the risk of re-identification decrease as the *k* value increases.

Table 2 presents an example of *k*-anonymity. In Table 2, note that each sequence of the values of four quasi-identifiers, Race, Birth, Gender, and ZIP occurs at least twice in the table.

Table 2 Example of *k*-anonymity, where *k*=2 and quasi-identifiers are Race, Birth, Gender, ZIP (Sweeny, 2002)

Race	Birth	Gender	ZIP	Problem
Black	1965	M	0214*	Short breath
Black	1965	M	0214*	Chest pain
Black	1965	F	0213*	Hypertension
Black	1965	F	0213*	Hypertension
Black	1964	F	0213*	Obesity
Black	1964	F	0213*	Chest pain
White	1964	M	0213*	Chest pain
White	1964	M	0213*	Obesity
White	1964	M	0213*	Short breath
White	1967	M	0213*	Chest pain
White	1967	M	0213*	Chest pain

2.4.2.2 *Spatial cloaking*

The *k*-anonymity framework can be directly applied to protection of location and trajectory privacy by considering that each location or trajectory as quasi-identifier. This approach is called 'spatial cloaking.' The basic idea of spatial cloaking is aggregating locations and trajectories into the spatial regions that contain at least *k* users for satisfying *k*-anonymity principle (Chow and Mokbel, 2012).

In the case of ATD, a preliminary spatial cloaking is implemented in the process of localization, because each user's location is approximated or aggregated to the location of the BTS tower to which the device is connected. Additionally, in the estimation procedure, the BTS locations are usually aggregated to larger spatial boundary systems such as traffic analysis zones (TAZs) and counties. Considering the size of the basic spatial unit and the huge sample size of ATD, one can expect that the possibility of re-identification from ATD is negligible (Ratti et al. 2006).

In the strict sense, however, the aggregation procedures involved in the OD demand estimation is not spatial cloaking, because they do not guarantee that the *k*-anonymity principle for location and trajectory privacy is satisfied. Also, the risk of re-identification may be increased when other quasi-identifiers exist in the dataset (e.g. time stamp, trip purpose, and trip pattern). Therefore, it is necessary for a data provider to satisfy *k*-anonymity in the dataset and to specify what *k* value is used for informing data users of the level of the risk of re-identification.

2.5 Discussion

This section reviewed the literature on the potential of ATD as a substitute for the traditional data collection tools for the OD demand estimation. Due to its cost-effectiveness and data adequacy, ATD has emerged a promising source for the OD demand estimation. Many scholars have proposed OD demand estimation methodologies for ATD collected by device tracking based on a communication network. The basic approach of the methods is to extract OD flow information from each user's trip trajectory and then to apply an adjustment factor to the initial information to scale it up to traffic flow.

There is, however, a definite need for validating the use of ATD before accepting it as a substitute for the traditional data collection tools. The three types of quality measures introduced in this review can be utilized for achieving this goal. First, absolute and relative error measures are particularly useful if a ground-truth table is available. Among the measures, the GEH statistic is a more engineering-oriented measure in that it places more emphasis on the percentage difference of larger flows. Also, the GEH statistic includes heuristically determined threshold values, making it convenient for interpretation. Secondly, in the absence of ground-truth data, one can use estimation error measures such as MPRE, ERE, and TDS when ATD is used as a source of link flow data to enhance and supplement to an OD demand estimation method that based on link flow counts. Since ERE takes into account all feasible OD demand matrices, it produces a more reliable measure of the deviation of the estimated table from the true but unknown table, compared to MPRE considering only the worst case of error. Thirdly, the statistical measure of 'the goodness of fit' such as R correlation and K-S statistic can be used for comparing the profiles of trip characteristics derived from two different OD demand matrices.

In our analysis below, the MAE and RMSE measures of fit are used due to their simplicity. Due to differences in sample size, the Pearson product-moment correlation coefficient is also used as a measure of relative fit. GEH values were estimated with the intention of measuring and mapping fit at the level of individual flows. However, this measure did not turn out to be useful due to differences in ATD sample sizes. GEH may prove more valuable after these data are scaled to more directly comparable population estimates.

Another important practical implication is that it is possible to re-identify an individual's location or trajectory from ATD even if records in ATD are anonymized and aggregated. Although the possibility may be extremely low, the use of ATD poses the risk of disclosure of location or trajectory information. Thus, data providers must pre-process ATD to satisfy *k*-anonymity before releasing the data and provide detailed information on privacy protection procedure they used.

2.6 Similar Research Efforts

2.6.1 NCHRP 08-95

The NCHRP 08-95 project, "Use of Cellular Data to Estimate Travel", is a parallel research effort being conducted concurrently by a different consulting team. In order to conserve resources, the literature review from NCHRP 08-95 is intended to complement this document when it is released. The focus of that effort is on the use of raw and processed passive mobile phone trace data.

2.6.2 FHWA Travel Model Improvement Program (TMIP)

In the spring of 2016, FHWA hosted a TMIP webinar focused on the use of archived travel data in support of transportation planning. The presentation was conducted by researchers from the Texas Transportation Institute (TTI) and covered their experiences over multiple projects. The following list highlights findings that are useful for guidance or exploration with the specific datasets in Ohio.

The following findings were identified for using cellular data for transport modeling:

1. Comparison on the whole matched well, but did not hold for smaller disaggregation areas
2. No vehicle type or mode
3. Not suited for small urban TAZs
4. 500x500 meter, as recommended by the vendor, is too small
5. Challenge for defining capture areas for EE trips
6. Misses short trips due to granularity of zones
7. Mixed results on trip purpose HBW trips underestimated due to trip chains
8. Commercial vehicle may be under-represented
9. EE trips were lower than ground counts
10. Applying a travel time constraint for EE trips is not possible
11. 300 meter accuracy for raw cellular data, 100 meter accuracy for aggregated activity locations
12. Assumption that the home is the overnight cluster and daytime is the work/school place
13. Trip types off of penetration rates and census demographics
14. External catch areas need to cover 45 minutes of drive time
15. Low correlation between employment estimates and estimated trips

The following findings were identified for using anonymous GPS data for transport modeling:

1. Anonymization of data does impact trip end accuracy
2. Sample penetration is increasing but still low and much lower than cellular data
3. Bias towards commercial vehicles
4. Routing information useful for model validation
5. Analyzed over long periods of time allows for better estimates of OD behavior
6. 10-20 meter accuracy for raw GPS
7. Commercial vehicle representation was very good
8. Possible to impute trip purposes due to positional accuracy

The following other findings may be useful

1. Time period aggregations are fine for all
2. Longer periods of data collection are better, one month is the minimum
3. Ideal purchase is a combination of products to meet specific needs

4. No known demographic biases, but plenty of suspicions and anecdotal evidence
5. Passive data has some “believability” with public and decision makers
6. Careful allocation of trips to stations based on AADT
7. Care with roads on TAZ or study area boundaries

2.7 Review of Applications of ATD in Transportation Planning

Transportation planning agencies have been increasingly considering ATD since the onset of Vendor A marketing in 2012. The following documented studies were evaluated in hopes of identifying findings and guidance related to data management and technical analysis. Additionally, project directors were contacted and interviewed regarding the effort to identify undocumented but useful information. It should be noted that several other states, regions, and cities are known to have purchased, evaluated, and integrated archived data products. These locations have not published assessments (or assessments could not be located) that contribute to the research objectives for this study.

2.7.1 South Alabama Regional Planning Commission (SARPC) – Mobile, AL – 2012

The NCDOT sent a short survey to Kevin Harrison, PTP at the SARPC that revealed the following details about their experience and success in using Vendor A data:

> **Has your department used mobile phone data (available from vendors such as Vendor A) for any of the above purposes? If so, please provide examples and any information on successful use of the data.**
>
> The Mobile MPO used Vendor A (with Alliance Transportation Group) to collect data for us to calibrate our gravity model. The cell phone data allowed us to create friction factors to calibrate the model by trip purpose. We did create an origin / destination matrix from just cell phone data. The mobile phone OD trip matrix was close to the modeled OD trip matrix; however, there were anomalies and what could you really do with it? What we were able to do, is produce a trip length frequency distribution curve and average trip lengths by purpose. Further, the study validated the NCHRP 2009 trip distribution percentages (see TRB presentation link), and made us realize that our trip generation was producing percentage of trips by purpose that were outdated. We have since updated them.
>
> In another study with Vendor A, we asked them to determine the home location of people using an 8 mile section of Interstate 10 crossing the Mobile Bay known as the “Bayway”. Since there could be minimal capture confusion (since it was a 8 mile bridge), and Vendor A archives data of a devices night (home) and day (work) location, we asked them to take a snapshot and tell us what state was the average “home” location of the devices that were on the Bayway on a couple of particular days. The data captured in a 24 hour period was only about a 10% sample size of ADT. That was beyond control of Vendor A, as that was the amount of cell users that were using their device on the Bayway. This helped us more accurately determine our External –External (EE) trip purpose; we were greatly underestimating the number of EE trips.
>
> **Has your department completed any comparison studies between mobile phone / GPS unit data collection versus traditional data collection activities?**
>
> The only real comparison I did was match the cell phone trip ends to the modeled trips ends by zone. This actually matched up closer than anticipated. That graph is in the TRB powerpoint link above. Interesting to note that the more rural zones had more trips ends in the cell phone data than what we are producing by our trip generation step in the model. Something we may consider investigating in the future (trip generation by land use).

2.7.2 Kentucky Transportation Cabinet – Kentucky – 2012-2014

The NCDOT sent a short survey to Jason Siwugla, PE at the Kentucky Transportation Cabinet and revealed the following details about their experience and success in using Vendor A data:

> Vendor A data was used during the development of the new Lexington model. In addition, Lexington Traffic Engineering has deployed approximately 35 "BlueTOAD" devices which read MAC addresses from Bluetooth enabled mobile devices and calculate travel speeds based on address matches at two or more locations. Speed reductions below normal threshold values trigger alterations in signal timing plans. The data is processed and archived by a vendor (TrafficCast International) but is available to the Lexington traffic management center in real-time. The use of archived BlueTOAD data for planning and congestion management purposes is being explored. The results are promising, but the devices have been operational for less than a year, so experience is somewhat limited.
>
> In addition to the MPOs mentioned, PB helped LAMPO do a model update and used Vendor A data. To date, we have not used them on a corridor type study in KY, either for KYTC Planning or another client.
>
> We do have experience with the data and the vendors in particular for used Vendor B data to calibrate speeds and travel time on I-70 in the KCMO area and will be using Vendor A data in the Nashville area commuter rail study I am heading up.
>
> The data must be put into the proper coding and sequence for use in the various models so someone with experience needs to do that.
>
> Usually the data can provide good inputs, especially for trip purposes and types that are typically may be hard to get at like special events (sports) or for colleges and universities or for Ft. Campbell in the case of the Nashville work.
>
> **Has your department used mobile phone data (available from vendors such as Vendor A or NavteQ/HERE) for any of the above purposes? If so, please provide examples and any information on successful use of the data**.
>
> (Vendor A) Development of Time of Day distribution of Trips for traditional travel demand models
>
> (Vendor A) Development of Trip Length Frequency Distributions by purpose and by Time of Day
>
> (HERE) Development of reference speeds for estimating free flow in models
>
> (HERE) Development of Time of Day average speeds for Corridors
>
> (NavteQ) Development of Time of Day speed profiles for Corridors
>
> **What problems or issues have been identified in relation to use of mobile data?**
>
> (Vendor A) Assignment of Trip Purpose can be skewed. You have to thoroughly review the purposes against know Zonal Landuse
>
> (Vendor A) Coverage can be spotty in rugged regions
>
> (HERE) Other data sources needed to vet "outlier" traveltime data
>
> (HERE) Data is rather aggregated and applies to "NHS" routes only
>
> (Navteq) Data is available beyond NHS at a LINK level, but can have 'holes'

2.7.3 North Carolina Department of Transportation (NCDOT) – Asheville, NC – 2013

The NCDOT and Parsons Brinckerhoff purchased Vendor A data to support travel demand model development for the French Broad River Metropolitan Planning Organization (FBRMPO). This ATD product was purchased to supplement the 2013 household travel survey (HTS), specifically to identify external travel (EE, EI, and IE). The project team cleaned and validated the data with the HTS. Data was purchased for one month and validated using the recent household travel survey and external station traffic counts. The study team found the data cost efficient, valid, and useful for its intended purpose. The team discovered that the Vendor A data included data for external travel between external zones that did not actually enter the study area. These trips had to be removed before being loaded into the model.

2.7.4 Napa County Transportation Planning Agency – Napa Valley, CA – 2014

Fehr and Peers conducted an assessment of travel behavior for Napa County, an area with a large number of visitors. This study made use of Vendor B Origin-Destination (TAZ to TAZ) Data to supplement traditional methods of intercept surveys, employment center surveys, license plate matching and traditional vehicle counts. By using an integrated approach, the Vendor B Data was refined to estimate personal vehicle trips by day of week and trip type. The trip type was derived from integrated results of a small survey with the large amount of data generated by Vendor B. This integration approach is a promising method for filling in known gaps in the travel details available from archived travel datasets.

2.7.5 Florida Department of Transportation – Northwest Florida – 2015

The Florida Model Task Force in partnership with the Florida DOT purchased archived data from Vendor A to support long range planning and model development for the Northwest Florida and Capital Region planning areas. The Vendor A data was for one month, included Mon-Thurs only, 24 hour totals, AM peak totals, and PM peak totals. The research team noted that the density of cell tower locations impacted the precision of the Vendor A estimates. A comparison between Vendor A estimates and three traffic count locations showed differences of approximately 20%. The team was able to load the data into the existing travel demand model and satisfactorily conduct select link analysis. It should be noted that this example allowed two adjacent model areas to be joined and processed using a single Vendor A dataset. This is not something that is easy to do using traditional survey methods.

2.7.6 West Contra Costa County – San Francisco – 2016

In a January 2016 report, The West Contra Costa County purchase Vendor A data for the San Francisco region to evaluate travel patterns in support of analysis regarding High Capacity Transit (HCT). The research team noted that the Vendor A data may be overestimating the number trips after comparison to existing model estimates. However, the relative distribution of trips between zones was very close to model estimates. Therefore the model team used the trip distribution percentages with other trip count data to generate total trips between zones. The team also noted that Vendor A compared very well for trips within the study area, but trips from outside the study showed higher variation from expected values. The reason for this difference was not identified but the team recommended close review and validation using additional data sources.

3 STUDY AREA AND BACKGROUND

3.1 Study area

The study area is Allen County: a rural county in northwestern Ohio with a population of 106,331 in 2010 (US Census Bureau 2010). Lima is the county seat. Major employers in Allen County include Ford Motor Company, General Dynamics, the Joint Systems Manufacturing Center, St. Rita's Medical Center, Husky Energy, and Procter & Gamble (https://development.ohio.gov). Overall, this is a rural area and consequently travel patterns will be relatively simple. In contrast, it is likely that corresponding travel patterns will be more complex in urban areas such as Cleveland, Columbus and Cincinnati. Therefore, the results in this study should be interpreted within this context; further study is required for travel patterns in more urbanized areas of Ohio.

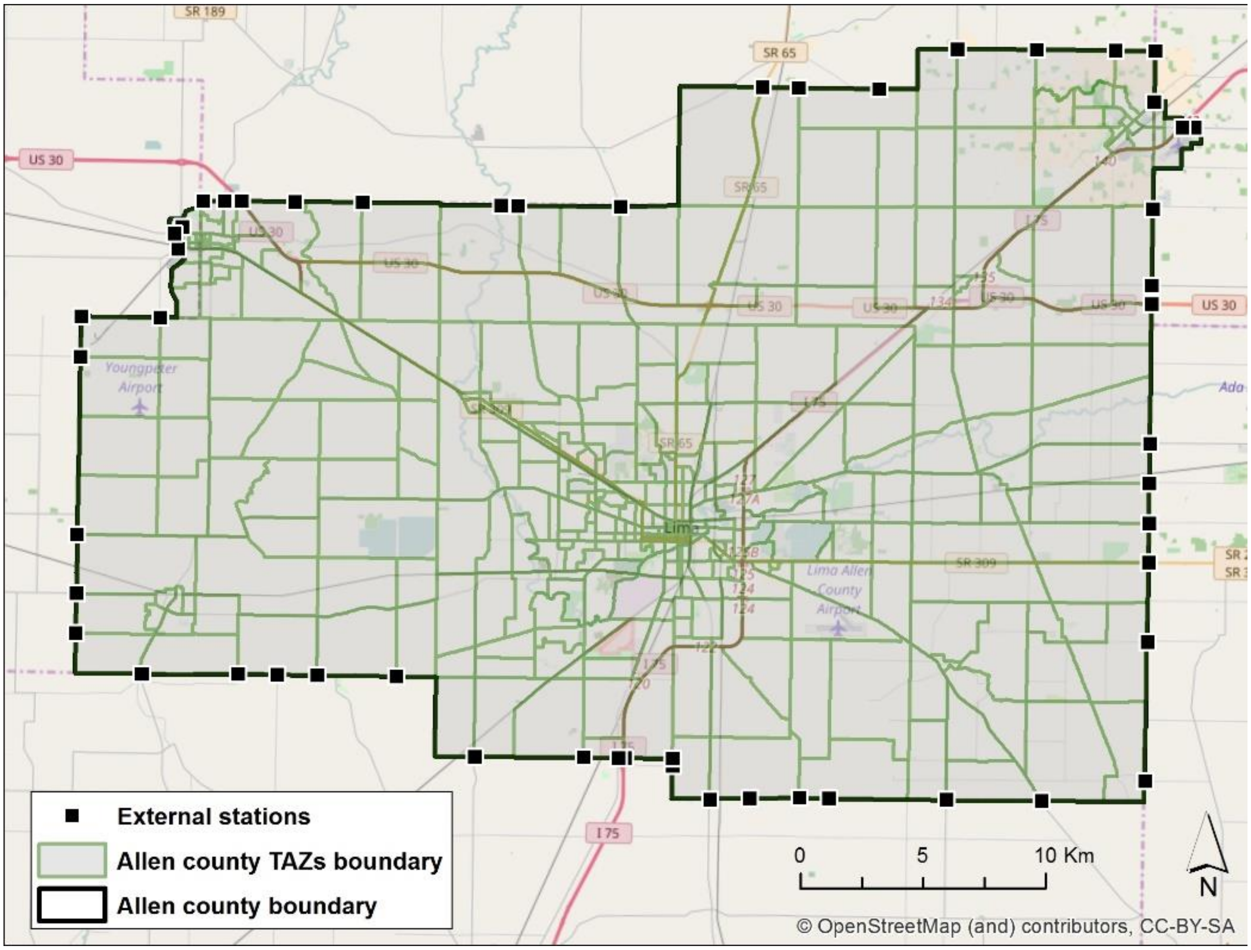


Figure 1: Allen County study area

3.2 Archived Travel Data Product Descriptions

ATD were acquired from three vendors; for purposes of this report, they are labeled as Vendor A, Vendor B and Vendor C:

- **Vendor A** data consists of generic vehicles (not distinguished, e.g., personal versus commercial vehicles). Vendor A derives their data entirely from cell phone signal based on triangulation from towers during phone activity.
- **Vendor B** data distinguishes between personal and commercial vehicles. This vendor derives data from navigation/traffic applications, extracting GPS tracks from users, as well as fleet/commercial vehicle GPS probe data.
- **Vendor C** data comprises GPS trajectories from commercial vehicles.

These data are described in more detail below.

3.2.1 Vendor A

The Vendor A data delivery included the following files:

1. ReadMe_AgeKey.docx – Descriptions of age group categories, see Appendix A
2. ReadMe_AutoKey.docx – Descriptions of auto ownership categories, see Appendix A
3. ReadMe_IncomeKey.docx – Descriptions of income distributions, see Appendix A
4. Readme_TripMatrixAttributes.pdf – Data dictionary for files, see Appendix A
5. trip_leg_matrix_cusWDDP.csv – Results for weekdays by time period groupings
6. trip_leg_matrix_cusWDH.csv – Results for weekdays by daily groupings
7. WDDP_age_matrix.csv – Results by age classification by time period
8. WDDP_income_matrix.csv – Results by income classification by time period
9. WDDP_veh_matrix.csv – Results by vehicle classification by time period
10. WDH_age_matrix.csv - Results by age classification by daily groupings
11. WDH_income_matrix.csv - Results by income classification by daily groupings
12. WDH_veh_matrix.csv – Results by vehicle classification by daily groupings

Figure 2 shows the zone structure for the data provided by Vendor A. The zones around the perimeter of the study area show the external catchment areas used by Vendor A to identify external trip ends and trip segments. These zones were defined by Vendor A. The green dots reference the 2008 Lima survey locations.

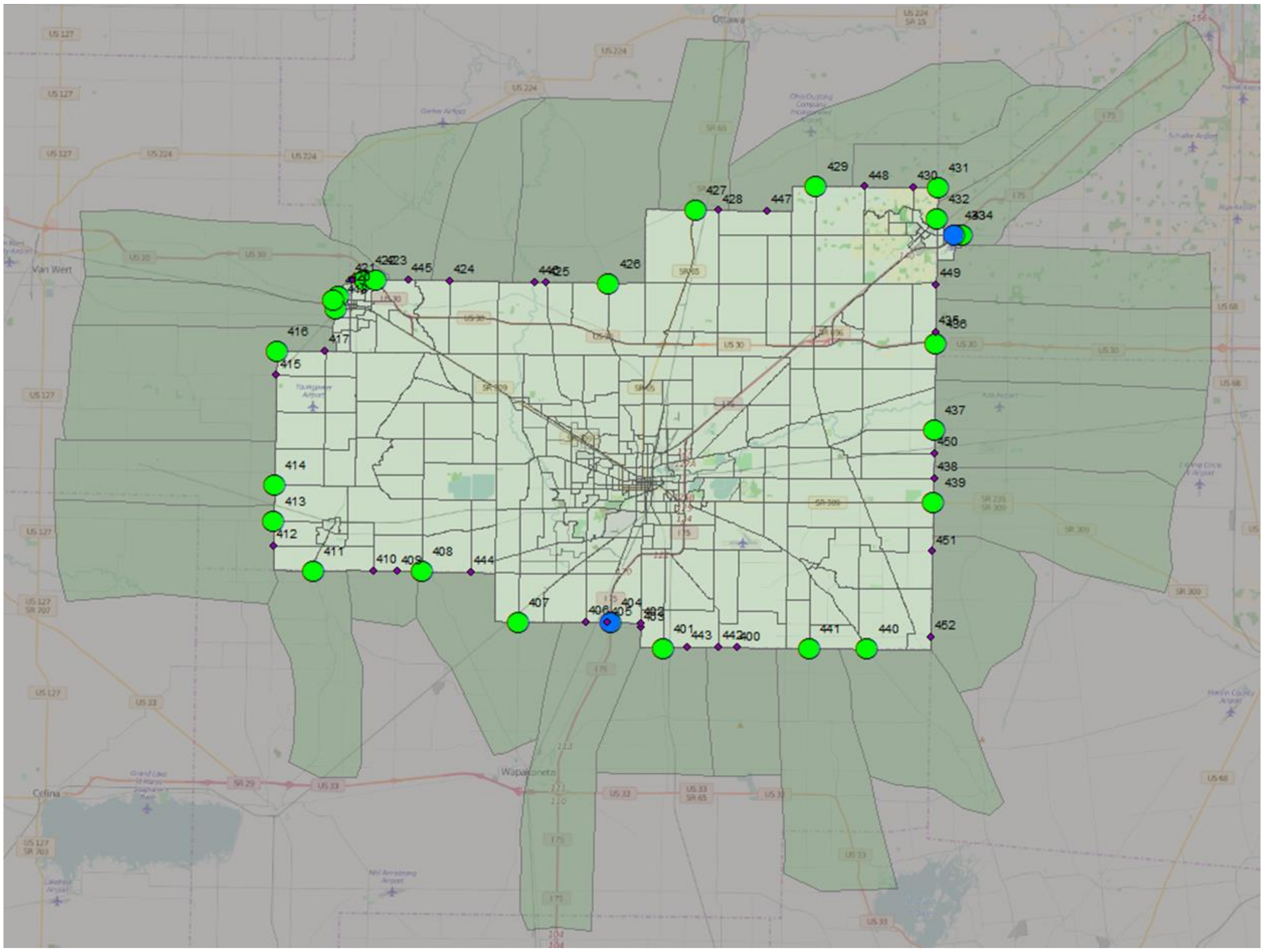


Figure 2 - Vendor A Zone Structure

3.2.2 Vendor B

The Vendor B delivery included the following files:

1. Destination_zone.shp: ESRI shapefile delineating zones for matching destinations
2. Lima_OD_7418_od_commercial.csv: Results for commercial vehicles
3. Lima_OD_7418_od_personal.csv: Results for personal vehicles
4. Lima_OD_7418_zone_frequencies_od_commercial.csv: Zone totals for commercial vehicles
5. Lima_OD_7418_zone_frequencies_od_personal.csv: Zone totals for personal vehicles
6. Lima_OD_7418_zones.csv: Table of zones
7. Origin_zone_set.shp: ESRI shapefile delineating zones for matching destinations
8. Project_OD.txt: Data dictionary (see Appendix A)
9. README-OD.txt: Further data descriptions regarding files, fields, and metrics

Figure 2 shows the zone structure for the data provided by Vendor B. The zones around the perimeter of the study area show the border zones areas used by Vendor B to identify external trip ends and trip segments. Vendor B actually uses narrow borders around the perimeter of the study to identify when vehicles enter/exit. Those vehicles are then either assigned a specific external station or a border crossing (occurred along a TAZ border but not at one of the 2008 survey stations). The green dots reference the 2008 Lima survey locations.

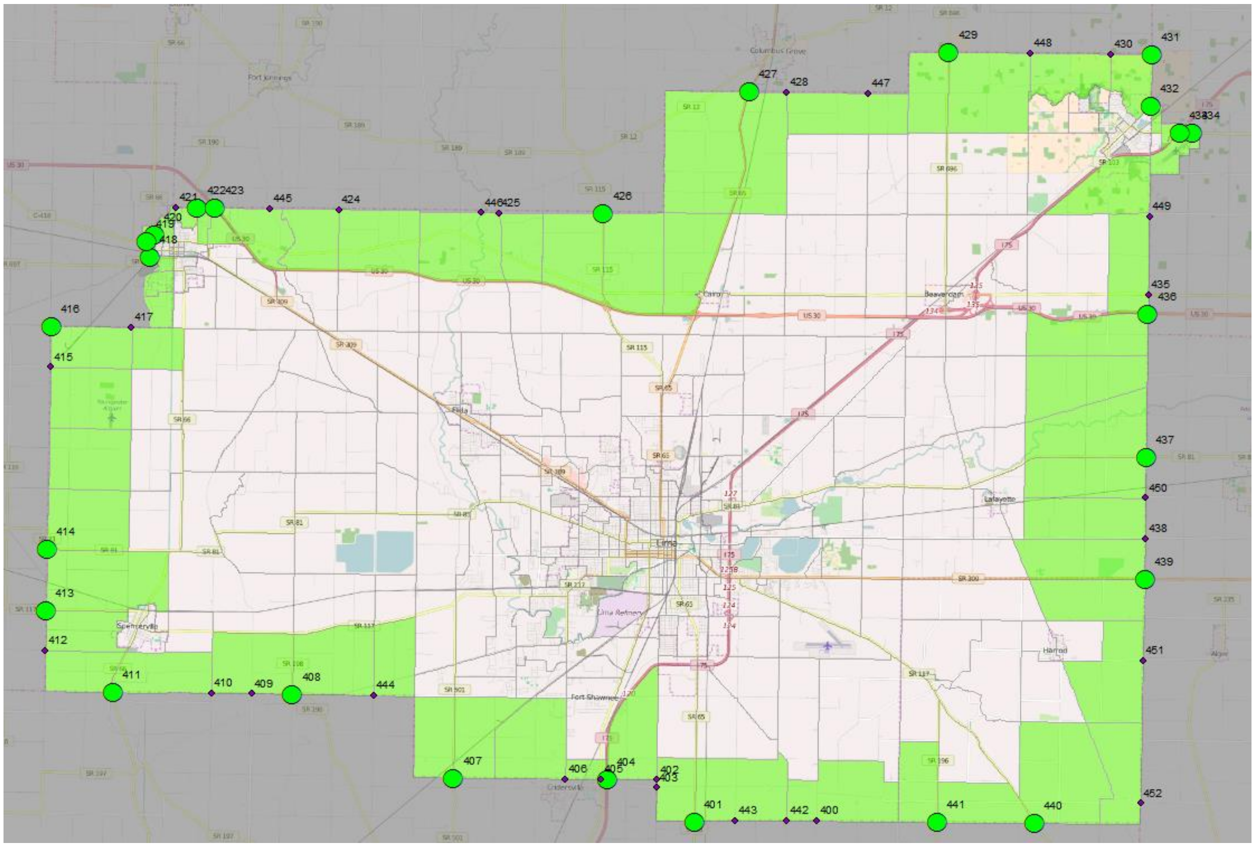


Figure 3 - Vendor B Zone Structure

3.2.3 Vendor C

The Vendor C delivery contained the following files:

1. Allen County Sequence June2015.csv – Result dataset
2. allenCountyDataDictionary.txt – Data dictionary (see Appendix A)

Figure 4 shows the zone structure used by Vendor C for the study. Vendor C used US Census Block Groups as locations in their data. Vendor C data do not include trip end information, only sequences of records by truck that includes a zone ID and a timestamp. The red highlighted block groups indicate what a trip may look like in the data: sequences of records as the vehicle moves from block group to block group. Note that the data must be processed into trip ends to generate estimated OD patterns.

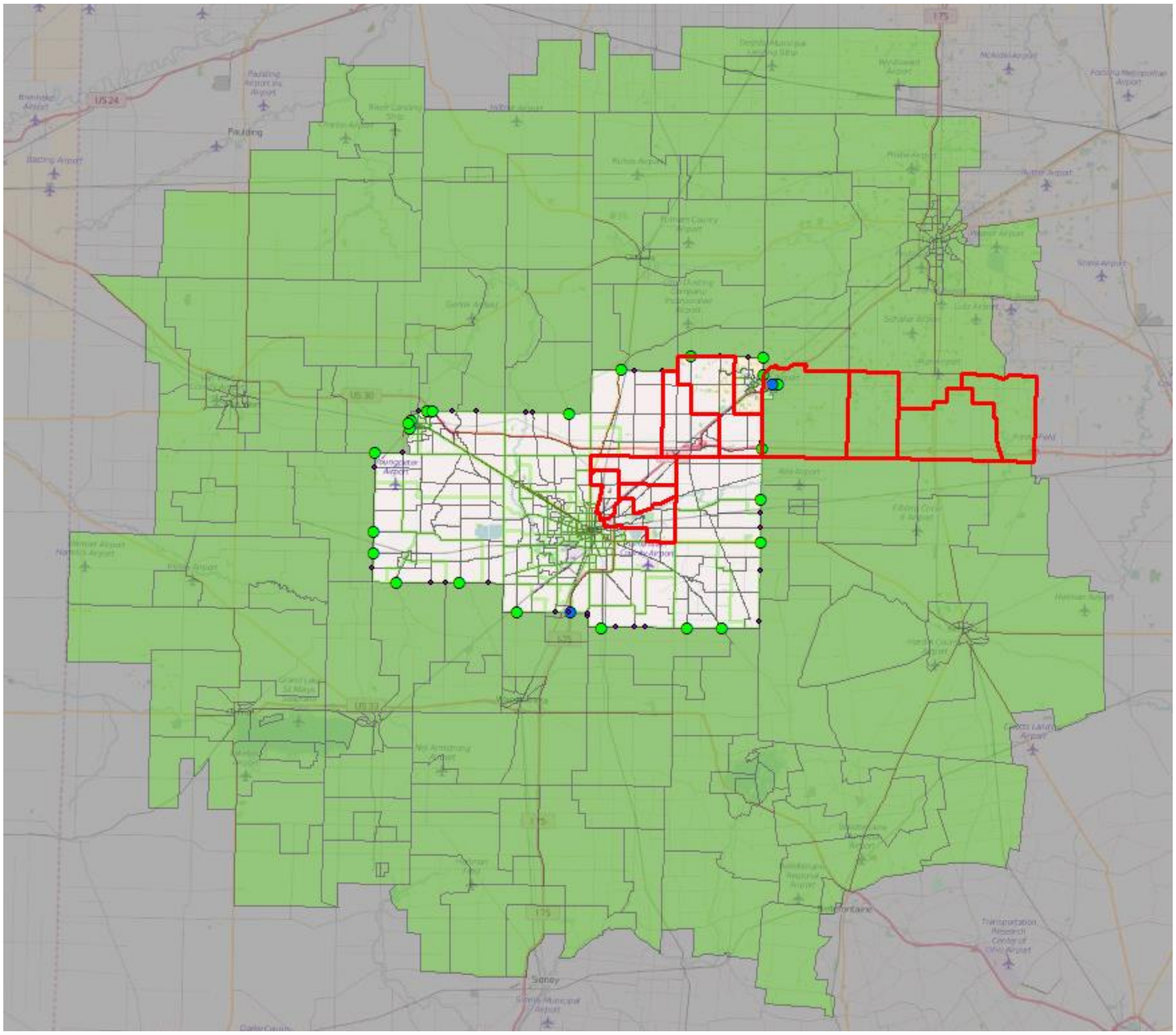


Figure 4 - Vendor C zone structure

3.3 Findings and Lessons Learned from Past Studies

Past studies and published research reports provide a foundation of experience and knowledge that can guide ODOT in its efforts to adopt ATD products into the external model development process.

3.3.1 Potential items to Investigate

1. For Vendor A, investigate whether internal zone size has an impact on relative trip distribution percentages and trip type percentages. Several studies noted the limitations of Vendor A data for small TAZs and in areas with sparse mobile phone coverage. Further, there is concern that trip type distributions based on location and time may result in false assumptions. If the Vendor A zone structure can be reduced in total numbers by aggregating TAZs, the total cost of the product is reduced. Further, it may be possible to improve upon TAZ trip distribution by disaggregating trips using local knowledge, demographics, and land use.
2. For Vendor B, investigate the personal vehicle trip distribution using past model estimates. There is concern by past users that the raw sample sizes of personal vehicle data available to Vendor B

is too small to capture the relative distribution of trips around a study area. While the identification of trip entry/exit points and trip ends can be accurate using GPS based data sources, the comprehensiveness of the data may lead to under-sampling in many areas, particularly in low volume or rural land uses.

3. External catchment areas for all products should be investigated. Past users investigating external travel have noted that the archived data product technology requires that moving vehicles be identified at their entry/exit points of the study area. Products based on GPS technology have an advantage in this regard due to the spatial resolution. However, Vendor C data, due to privacy protection methods, obfuscates their data such that the entry/exit points are in question. Vendor A typically defines external catchment areas as another layer of zones and considers them computationally similar to TAZs. Therefore, trips originating beyond those TAZs may not be identified as their identification relies entirely on finding a moving vehicle as opposed to a trip end.
4. The amount of archived data to achieve the best results should be investigated. Past purchases have typically been made for one month of data. However, there is evidence that longer durations may achieve better results. While this also likely increases cost, there should be a known trade-off with data quality and this knowledge does not currently exist. In fact, it may be that the quantity of data may vary by location (urban areas with higher density require shorter durations than rural areas with lower density).

3.3.2 Validation of Data

1. Comparison of relative zonal distributions. This was shown to be more effective than using ground counts because the actual number of trips estimated by archived data is known to have flaws. This can be mitigated by using only trip distribution percentages between external stations and TAZs and factoring these values by the actual ground counts.
2. Comparison of trip type distributions by TAZ. It has been suggested that there is a poor mismatch between trip type estimates and expected values. The approach used by the archived data vendors to estimate trip type is likely something that can be improved with better algorithms. Therefore, the comparison of trip type distributions can identify if the archived data needs to be re-processed or mitigated in some manner.

3.3.3 Data Cleaning

1. External to External trips. Multiple investigators noted that Vendor A had a high number of trips between external stations that occur in adjacent catchment areas. These trips likely never traveled into the study area but simply between the catchments areas outside of the study areas. These trips should be cleaned.
2. Assigning of trips to external stations. Most efforts that discussed this topic assigned external trips from a catchment area to external stations based on AADT. It was noted in one case that this can result in problems in that trips can be mis-assigned to their entry/exit points resulting in incorrect route assignments. A better method is to assign trips to entry/exit points based on the trip ends and shortest time path.

4 ANALYSIS OF ARCHIVED TRAVEL DATA

4.1 Methodology

The main purpose of this study is to explore the potential of ATD as a substitute for existing travel data collected by traditional roadside survey through the comparison of trip counts of archived travel data (ATD) with ODOT trip tables. This section discusses the data preparation of the ODOT and ATD datasets.

4.1.1 ODOT data

ODOT provided network data and the results of a road-side survey and a camera license plate survey conducted in 2008 and 2009 and two fully-populated trip tables for 2006 and 2011 built based on the survey result.

The ODOT highway network layer was imported to TransCAD from a shape file. The highway network is built from this layer and includes all the associated data from ODOT. The speed field was checked and a few records with missing data were discovered; these were all highway ramps. These missing data were replaced with a nominal speed of 25 mph. These are very short segments and should not affect the analyses.

The network was built with the centers of the TAZ and external TAZ nodes with numeric IDs corresponding to the 1 to 452 range; external TAZs have IDs of 400 or greater. The TAZ centroids were obtained from: i) internal parcel file, and; ii) an external station file. Since a point location was needed to represent TAZs, a simple analysis of the parcel file was performed to determine average of the parcel locations inside the same TAZ, designating this as the point location representing that TAZ. The TAZ centers were tagged with the ID of the nearest network node.

Travel time estimates for each network segment were calculated using the formula:

$$\text{TIME} = \text{DIST}\ (60/\text{SPD})$$

where DIST is length in miles and SPD is the posted speed limit in miles per hour. Due to the lack of reliable data, this was not modified with turn penalties based on road hierarchy / classification. These speed estimates served as the basis for constructing a quickest path matrix between every pair of nodes. The data were checked for validity using straight line distances between all the TAZ nodes. For example, the node pair (401, 429) had a straight line distance of approximately 20 miles and an estimated travel time of approximately 23 minutes. This is consistent with the actual geography of the county. Note that the actual drive time is variable and this estimate is for uncongested speeds with zero delay or turn penalty.

4.1.2 Preparing Archived Travel Data (ATD) for Analysis

All original ATD products were imported into the same OD table structure to ease analysis. The table structure included the following fields:

Table 2 - Raw data fields used for analysis

Field Name	Data Type	Description
ID	Number	1000000-Vendor A, 2000000-Vendor B, 3000000-2008 Survey, 4000000-Vendor C, 5000000-model
From Zone	Text	Origin Zone Name in the original data
From Zone Type	Text	Internal or External
To Zone	Text	Destination Zone Name in the original data

To Zone Type	Text	Internal or External
Source	Text	"Vendor A", "Vendor B", "Vendor C", "2008Survey", "Model"
Count	Number	Number of trips
DayType	Text	"Weekday", "Weekend"
TimePeriod	Text	"Daily", "AM", "MD", "PM", "NT"
VehicleType	Text	"Commercial", "Personal", "Unknown"
TripPurpose	Text	"HH","HO","HW","OO","OH","OW","WW","WO","WH","B","P","W"
ExtTripType	Text	"IE", "EI", "EE", "II"
PersonType	Text	varies based on Vendor A terms

Vendor A data did not identify specific external stations for their trip entry/exit points. Instead, they used catchment areas (see Figure 2). The catchment areas could encompass one or more external station locations. In situations where one or more external stations existed, an ADT weighted assignment process was used to estimate the entry/exit point for trips. Results of analysis showed large numbers of short trips entering and exiting the study area and traveling to adjacent catchment areas. It was determined that these trips likely never entered the study area. Further, short IE or EI trips showed odd estimated entry/exit points given the origin and destination zones. A new approach of assigning entry / exit points was devised that assigned a centroid location for each catchment zone and network links that followed existing road paths to the study area as well as adjacent zones. Trips were then assessed using a shortest time path between the origin and destination zone centroids. Estimated entry / exit points were then assigned based on this analysis.

Vendor B data identified external station entry and exit points for the trip data and no additional entry/exit point estimation was needed.

Vendor C data were provided in a raw format that was processed from original GPS data but delivered as a set of vehicle locations and timestamps where the location was presented as a US Census Block group (no GPS data provided). Further, vehicle IDs were regenerated at regular intervals thereby preventing the identification of a single vehicle's full month activity pattern. To estimate the entry / exit point for trips, a shortest path approach was applied between the origin zone and the destination zone with an additional path requirement of following network links within or passing through the original data's US Census block groups. It was determined that the US Block Group definition was too coarse for effective OD analysis and Vendor C provide TAZs for travel within the study area.

4.1.3 Comparison of the datasets

Table 3 summarizes the properties of the ODOT standard data and the ATD from the three vendors. Trip type refers to the type of flow relative to the cordon: II is internal-internal trips (do not cross the cordon boundary); EI/IE are flows that cross the cordon boundary – external to internal and internal to external trips, respectively. Vehicle type classifies vehicles to personal, commercial or truck, depending on the data source. Finally, trip purpose refers to a classification based on survey data (in the case of ODOT) or trip ends (in the case of Vendor A).

Table 3: Overview of datasets

Dataset	Trip type II	 EI/IE	 EE	Vehicle type	Trip purpose
Unexpanded 2008 Roadside survey result	-	Y	Y	Commercial/ Personal	Work/Non-work/Truck
Expanded 2008 Roadside survey result	-	Y	Y	-	Work/Non-work/Truck
2008 Roadside survey and 2009 ALRP on IR 75	-	-	Y	-	-
2006 model external trip table	-	Y	Y	-	Work/Non-work/Truck
2011 model external trip table	-	Y	Y	-	Work/Non-work/Truck
Vendor A: Weighted	Y	Y	Y	Unknown	HH, HO, HW, OH, OO, OW, WH, WO, WW
Vendor A: Shortest Path	Y	Y	Y	Unknown	HH, HO, HW, OH, OO, OW, WH, WO, WW
Vendor B: Personal	Y	Y	Y	Personal	Unknown
Vendor B: Commercial	Y	Y	Y	Commercial	Unknown
Vendor C	Y	Y	Y	Truck	Truck
Notes: **i) Trip type: II is internal-internal; EI is external to internal; IE is internal to external; EE is external to external. ii) Trip purpose: H is home; W is work; O is other.**					

Table 3 shows that, although Vendor A data include all types of vehicles, Vendor B had data differentiated as commercial and personal vehicles, while Vendor C has data on trucks. Since each ATD has different types of vehicles, only direct comparisons among selected datasets could be made. Table 4 provides the data comparisons conducted in this analysis. The personal and type vehicles in Vendor B data were assumed to correspond to Work/Non-work purpose and Truck vehicles in ODOT data, respectively. The ODOT data for comparison is the most recent model input external trip table, the 2011 model external trip table.

Table 4: Data comparisons

ODOT data	Archived travel data
2011 model external trip table - all	Vendor A: WTD
2011 model external trip table - all	Vendor A: SP
2011 model external trip table – Work/Non-work	Vendor B: Personal
2011 model external trip table – Truck	Vendor B: Commercial
2011 model external trip table – Truck	Vendor C

4.2 Results

4.2.1 Trip length

Trip lengths (with respect to travel time) using the estimated network travel times as described above were computed. These data were analyzed using a procedure that categorized trips into one of six intervals: 1) less than 5 minutes; 2) greater than or equal to 5 minutes and less than 10 minutes; 3) greater than or equal to 10 minutes and less than 15 minutes; 4) greater than or equal to 15 minutes and less than 20 minutes; 5) greater than or equal to 20 minutes and less than 25 minutes; 6) greater than or equal to 25 minutes.

Table 5 provides trip length distribution estimates for external-external flows based on the different data sets in the analysis. These counts are useful for interpreting the absolute goodness of fit measures (absolute mean error, root mean square error) provided in Table 6. Mean absolute error is the average of the absolute errors while root mean square error is the sample standard deviation of the differences between the corresponding data. Both measures are scale-dependent and are therefore sensitive to differences in sample size. We also report the square of the Pearson product-moment correlation coefficient ("R-squared") as a measure of relative fit. (Recall that EE flows less than five minutes in length were eliminated as error.)

As mentioned above, the datasets in the analysis vary widely with respect to sample size. This is apparent from the estimates in Table 5 – in particular, Vendor B's data had much larger data values than the other vendors and ODOT data. This is assumed to be the result of the scaling of raw observations and not due to larger sample sizes. Regardless, the values provided to the research team were evaluated without any assumptions and evaluated using absolute values and relative comparisons. The absolute error measures in Table 6 reflect these differences in data scales: note the very large absolute error measures for Vendor B's data. However, the R-square measure shows better fit since this is a scale-independent measure of relative fit. Values close to 1.00 for this measure indicate that the relative pattern in the given ATD fit the relative pattern in the ODOT well. Table 6 suggest that all of the vendor data reproduces the relative patterns of EE trip length in the corresponding ODOT data, with the exception of Vendor A with the AADT-based weighted assignment method, the relative fit of which is poor.

Table 5: EE trip length distribution estimates

EE trip length	< 5	5-10	10-15	15-20	20-25	> 25	Sum
ODOT 2011	0	243	374	1502	24023	1702	27,843
Non-work	0	133	206	852	12061	1079	14,331
Work	0	71	69	306	1518	166	2130
Truck	0	39	99	343	10444	457	11,383
Vendor A: WTD	0	8361	4046	976	1763	334	15,479
Vendor A: SP	0	89	124	204	1583	131	2131
Vendor B: Personal	0	8890	3085	7290	132,068	12,284	163,617
Vendor B: Commercial	0	105,243	7064	35,712	1,076,545	72,312	1,296,875
Vendor C	0	105	44	209	20,986	2401	23,745
Note: **EE trip lengths of less than 5 minutes have been eliminated as error**							

Table 6: EE trip length distribution - overall goodness of fit

ODOT data	ATD	Absolute mean error	Root mean square error	R Square
ODOT 2011 (total)	Vendor A: WTD	7188.9	10,743.0	0.081
ODOT 2011 (total)	Vendor A: SP	5142.4	10,077.6	0.997
ODOT 2011 Non-work + work	Vendor B: Personal	29,431.3	53,446.0	0.995
ODOT 2011 Truck	Vendor B: Commercial	257,098.5	480,437.8	0.992
ODOT 2011 Truck	Vendor C	2548.3	4794.5	0.993

The following figures provide more detail depictions of the fit between ATD estimates of EE trip length relative to ODOT data. Figure 5 compares data for all vehicles from ODOT 2011 versus Vendor A data; WTD refers to the AADT-based weighting methods for assigning flows to external stations, and SP refers to the shortest path method. Figure 6 compares data about personal vehicles from ODOT 2011 versus Vendor B. Figure 7 compares ODOT 2011 truck data with similar data from Vendor B and Vendor C. These figures confirm the overall goodness of fit results in Table 6: the ATD-based EE trip length estimates match the corresponding ODOT data well, with the exception of Vendor A: WTD which greatly over estimates trip length counts in the 5-10 minute and 10-15 minute categories.

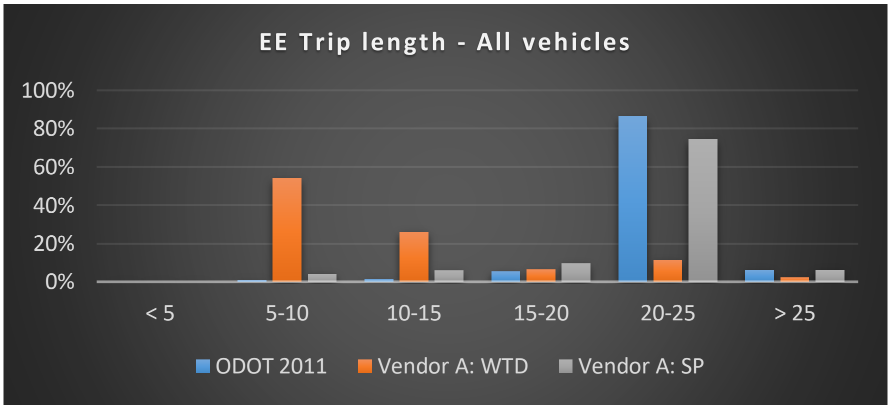


Figure 5: EE trip length comparison for all vehicles- ODOT 2011 vs. Vendor A

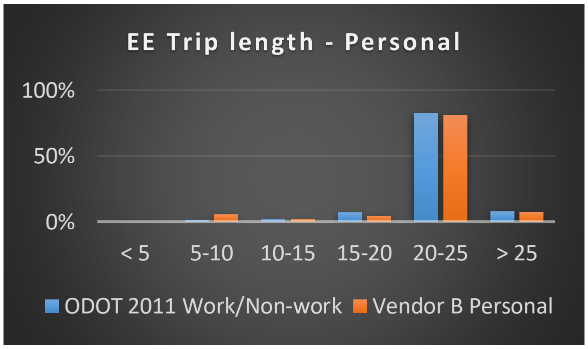


Figure 6: EE trip length comparison for personal vehicles – ODOT 2011 vs. Vendor B

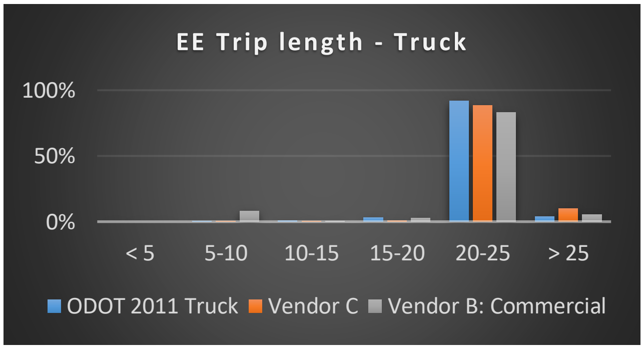


Figure 7: EE trip length comparison for trucks – ODOT 2011 vs. Vendor B and Vendor C

Table 7 provides estimates of trip lengths for external-internal/internal-external flows and Table 8 provides overall goodness of fit for the ATD with the ODOT data. As Table 8 suggest, the ATD-based estimates of EI/IE trip lengths do not fit the corresponding ODOT data well, with the exception of Vendor B data on personal vehicles and Vendor C data on trucks. But, even in these cases, the relative fit is not as strong as with EE trip length estimates.

Table 7: EI/IE trip length distribution

EI/IE trip length	< 5	5-10	10-15	15-20	20-25	> 25	Sum
ODOT 2011	26,863	19,656	21,147	17,579	9649	3280	98,173
Non-work	15,609	9957	12,308	9868	5374	1249	54,365
Work	9087	7528	7602	6554	3182	707	34,660
Truck	2166	2171	1238	1157	1093	1323	9148
Vendor A: WTD	4932	10,219	10,018	8725	4073	509	38,477

Vendor A: SP	5858	10613	9937	8430	3820	479	39,137
Vendor B: Personal	38,068	54,949	47,775	37931	19,203	3548	201,473
Vendor B: Com	106,318	253,844	212,379	288,226	76,263	8226	945,254
Vendor C	10,745	10,623	3915	4180	580	79	30,122

Table 8: EI/IE trip length - overall goodness of fit

ODOT data	ATD	Absolute mean error	Root mean square error	R Square
ODOT 2011 (total)	Vendor A: WTD	9949.4	11,626.3	0.444
ODOT 2011 (total)	Vendor A: SP	9839.4	11,359.0	0.543
ODOT 2011 Non-work + work	Vendor B: Personal	18,741.3	22,125.9	0.707
ODOT 2011 Truck	Vendor B: Commercial	156,017.6	185,685.1	0.012
ODOT 2011 Truck	Vendor C	4081.3	5214.5	0.836

The following figures provide more detail depictions of the fit between ATD estimates of IE/EI trip length relative to ODOT data. Figure 8 compares IE/EI trip lengths for all vehicle using data from ODOT 2011 versus Vendor A, again using both methods for assigning flows to external stations. Figure 9 compares trip lengths for personal vehicle using data from ODOT 2011 versus Vendor B. Figure 10 compares trip lengths for trucks using data from ODOT 2011 and both Vendor B and Vendor C. As suggested by the poor to mediocre goodness of fit measures, there are noticeable qualitative differences in the trip length patterns between the ATD and the corresponding ODOT data.

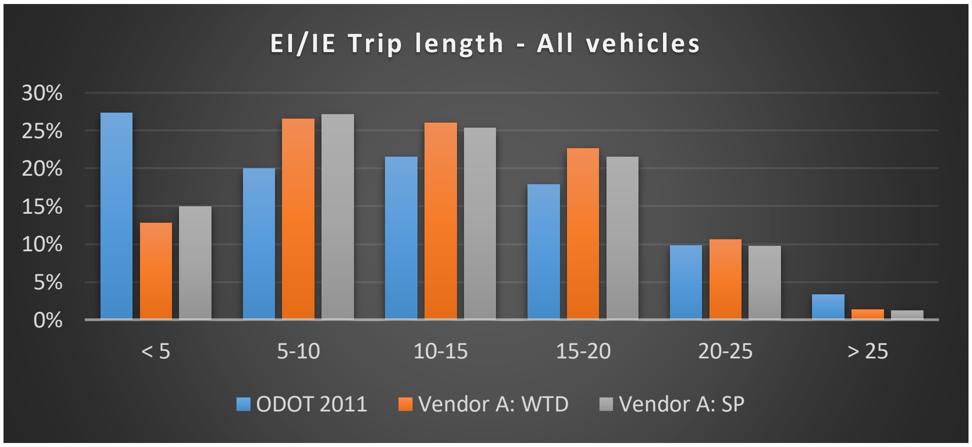
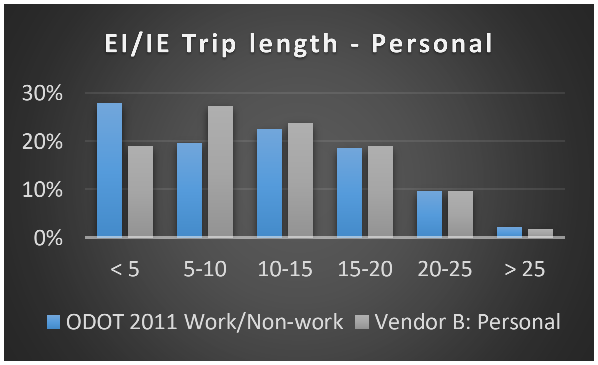


Figure 8: EI/IE trip length comparison for all vehicles – ODOT 2011 vs. Vendor A



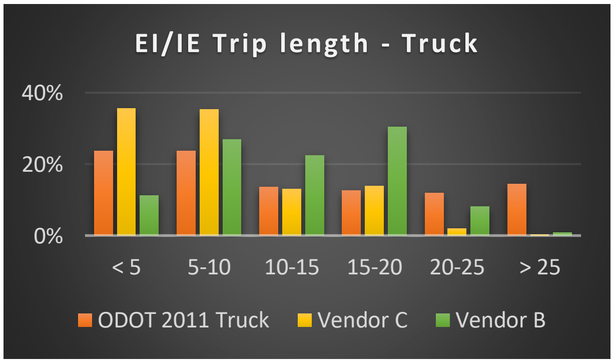


Figure 9: EI/IE trip length comparison for personal vehicles – ODOT 2011 vs. Vendor B

Figure 10: EI/IE trip length comparison for trucks – ODOT 2011 vs. Vendor B and Vendor C

Per comments by ODOT on June 8 2017 on the initial draft of this report, some of the trip length analysis was repeated at a higher level of temporal resolution, specifically, using one minute bins. **Figure 11**, **Figure 12** and **Figure 13** show comparisons for EE flows between ATD and ODOT data for all vehicles, personal vehicles and trucks (respectively) using one-minute bins; these correspond to Figure 5, Figure 6 and Figure 7 above. These results mirror the results above based on 5 minute bins, albeit with greater detail.

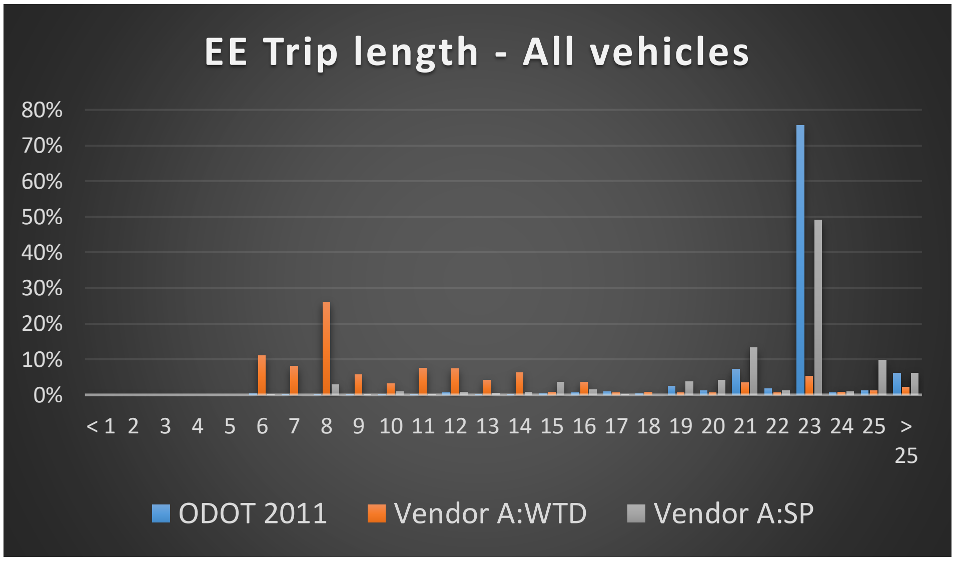


Figure 11: EE trip length comparison for all vehicles- ODOT 2011 vs. Vendor A – one-minute bins

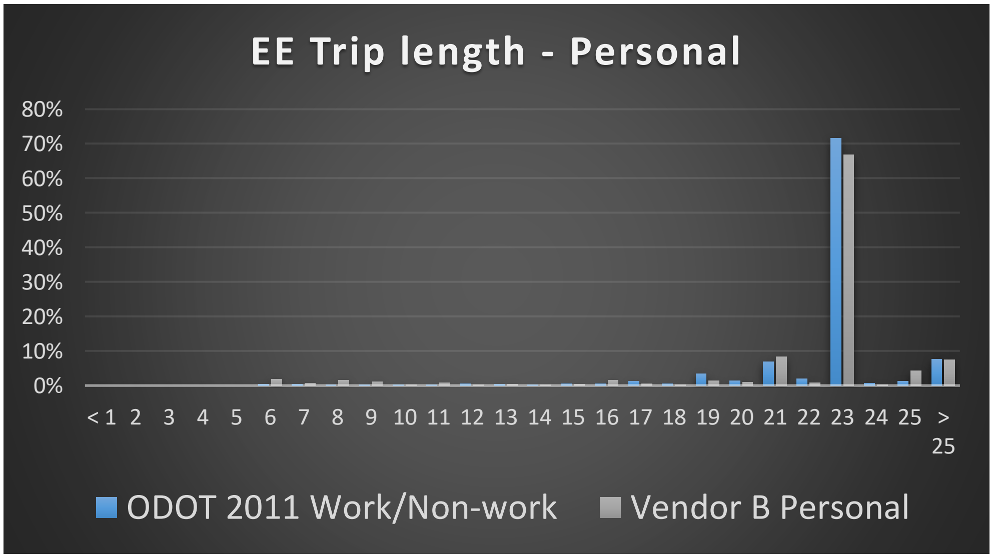


Figure 12: EE trip length comparison for personal vehicles – ODOT 2011 vs. Vendor B – one minute bins

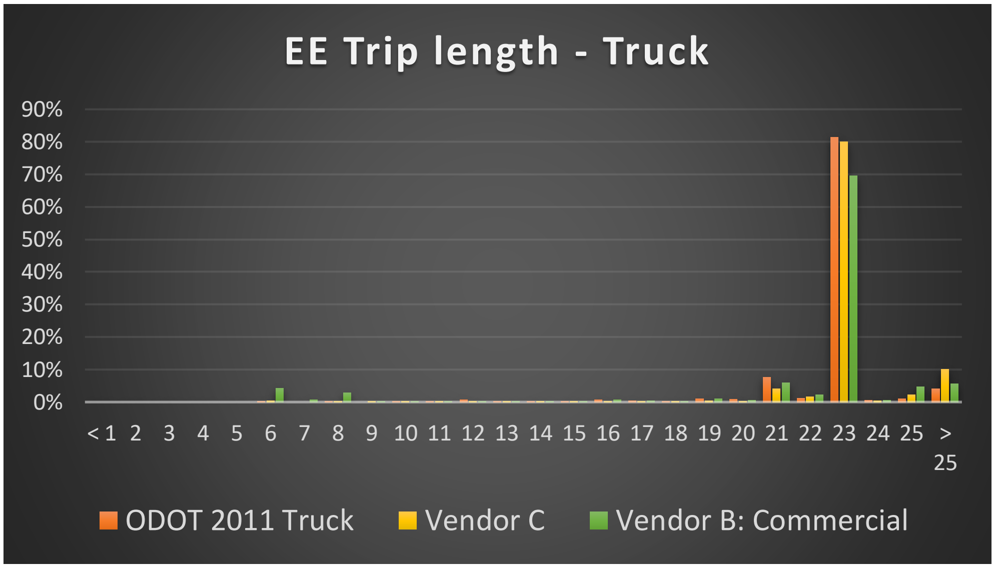


Figure 13: EE trip length comparison for trucks – ODOT 2011 vs. Vendor B and Vendor C – one minute bins

Figure 14, Figure 15 and Figure 16 provide the one-minute bin analysis for EI/IE trip lengths for all vehicles, personal vehicles and trucks, respectively; these figures correspond to Figure 8, Figure 9 and Figure 10. Again, these results mirror the results above based on 5 minute bins, albeit with greater detail.

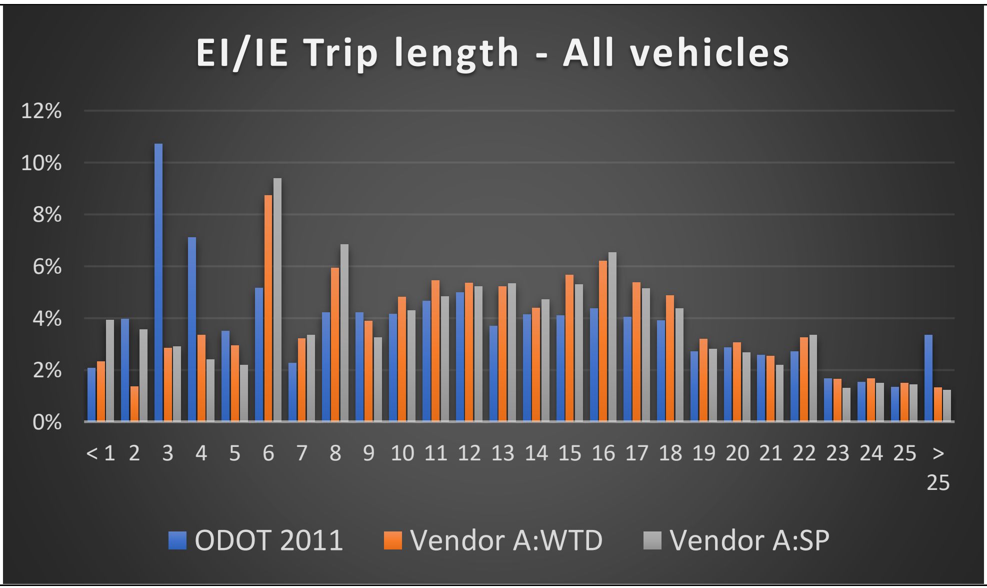


Figure 14: EI/IE trip length comparison for all vehicles – ODOT 2011 vs. Vendor A – one minute bins

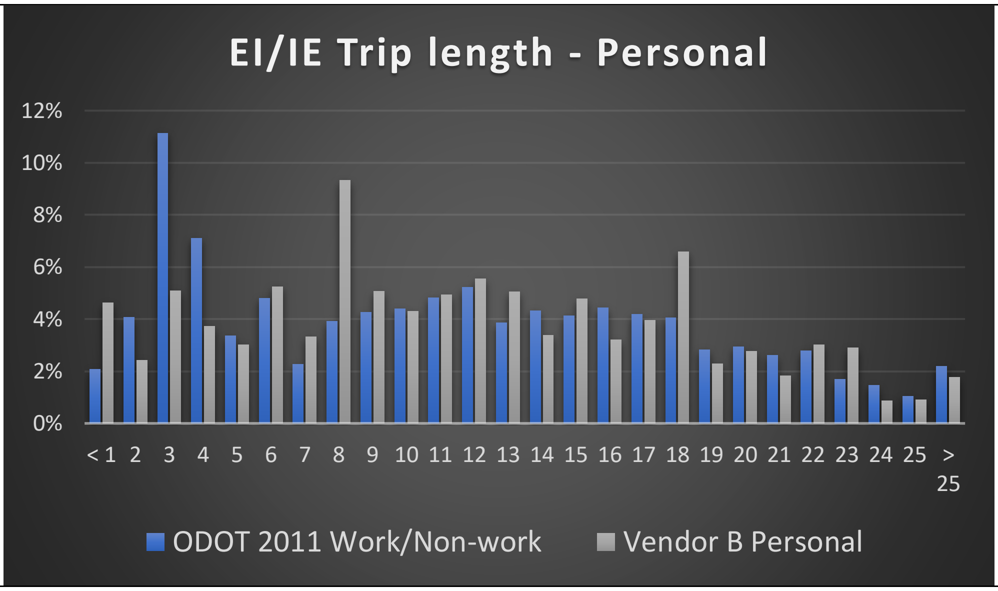


Figure 15: EI/IE trip length comparison for personal vehicles – ODOT 2011 vs. Vendor B – one minute bins

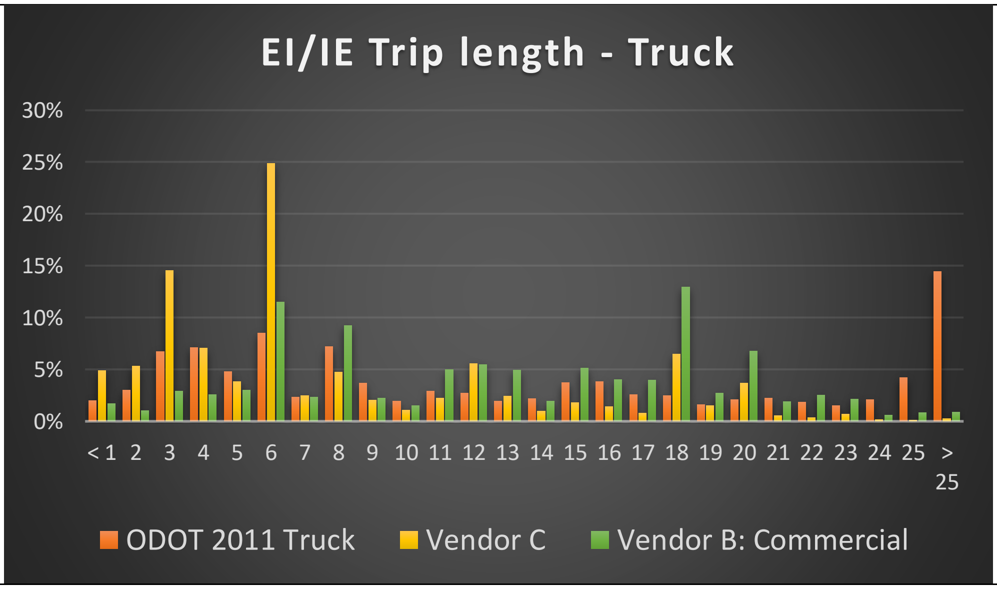


Figure 16: EI/IE trip length comparison for trucks – ODOT 2011 vs. Vendor B and Vendor C – one minute bins

Appendix 7.8 of this report contains data tables corresponding to the one-minute bin analysis for both EE and EI/IE flows, expressed as percentages. These data can be used for further analysis.

4.2.2 Trip purpose

The trip type analysis of the archived data products evaluated the ability of a product to match the business, personal, and commuting trip patterns observed in the 2008 external cordon survey. Vendor A is the only company that provided estimates of trip types based on trip ends. Table 9 provides the reconciliation between Vendor A trip purposes based on trip ends and the ODOT trip purpose designation.

Table 9: Reconciling trip purposes in Vendor A and ODOT data

Vendor A trip purpose	ODOT trip purpose
HH: Home to Home	Personal
HO: Home to Other	Personal
HW: Home to Work	Commute
OH: Other to Home	Personal
OO: Other to Other	Personal
OW: Other to Work	Business
WH: Work to Home	Commute
WO: Work to Other	Business
WW: Work to Work	Business

In a simple comparison between the raw survey data and original Vendor A data (external to external (EE) and external to internal (EI) trips only), there are substantial differences (see Table 10) even before specific external station or zonal comparisons. Vendor A shows a much larger percentage of personal trips than were observed in the 2008 ODOT survey data.

Table 10: Overall external trip type comparison between Vendor A and ODOT survey

	Business	Personal	Commute
Vendor A	7.9%	84.3%	7.7%
ODOT 2008 Survey	18.7%	48.2%	33.1%

Vendor A also provides a “person type” using six different categories as shown in Table 11. In that table, “Vendor A: WTD” refers to the dataset that was allocated to external stations based on station traffic volumes. “Vendor A: SP” refers to the dataset allocated using shortest network path methods. In both cases, the inbound/outbound commuter type percentage is higher than the estimated EE and EI commute trip percentages. The person type and trip types should not match exactly (trips of various types can be taken by all person types).

Table 11: Vendor A person type distribution

	Vendor A: WTD	Vendor A: SP
Long Term Visitor	8.1%	9.6%
Short Term Visitor	69.6%	54.8%
Inbound Commuter	10.5%	14.8%
Outbound Commuter	4.1%	6.7%
Resident Worker	5.2%	9.2%
Home Worker	2.5%	4.9%

We used the Vendor A - SP dataset for further comparisons by ingress external station as well as the destination traffic analysis zone (TAZ) or external station. Table 12 lists the comparison by the ingress external station for the 2008 ODOT Survey and Vendor A trip percentages by business, personal, and commute trip types. The differences indicate very large error ranges even at this aggregate level. Figure 17 is a scatter plot of the Vendor A and survey commute trip percentages. There is no discernable pattern in the scatter plot that suggests a relationship between the variables.

Table 12: Trip type comparisons by ingress external station

	Business		Personal		Commute	
External Station	ODOT	Vendor A	ODOT	Vendor A	ODOT	Vendor A
401	11.8%	1.4%	40.9%	91.6%	47.4%	7.1%
407	10.9%	1.4%	50.7%	66.9%	38.4%	31.8%
408	10.4%	0.0%	60.8%	65.4%	28.8%	34.6%
411	19.4%	3.0%	45.6%	66.7%	35.0%	30.3%
413	17.2%	0.0%	39.6%	0.0%	43.2%	0.0%
414	17.2%	7.8%	40.4%	84.4%	42.4%	7.8%
416	9.8%	7.5%	47.1%	82.6%	43.1%	9.9%
418	14.2%	6.0%	48.0%	78.5%	37.8%	15.5%
419	21.2%	0.0%	48.5%	98.6%	30.3%	1.4%
420	20.1%	0.0%	50.8%	98.6%	29.1%	1.4%
422	16.2%	0.0%	50.4%	100.0%	33.4%	0.0%
423	31.5%	1.0%	42.9%	92.9%	25.6%	6.1%
426	14.0%	7.4%	40.5%	75.6%	45.6%	17.0%
427	18.9%	4.6%	42.2%	74.7%	38.9%	20.6%
429	18.9%	9.0%	48.8%	68.5%	32.3%	22.5%
431	15.8%	4.2%	47.1%	91.6%	37.1%	4.2%
432	11.0%	0.0%	47.7%	100.0%	41.3%	0.0%
434	11.9%	0.0%	54.4%	100.0%	33.7%	0.0%
436	37.4%	3.7%	46.1%	79.1%	16.5%	17.3%
437	10.9%	7.0%	43.9%	72.2%	45.2%	20.7%
439	14.3%	10.8%	58.3%	74.7%	27.5%	14.5%
440	27.0%	6.3%	51.7%	84.1%	21.3%	9.6%
441	7.4%	5.3%	62.8%	79.7%	29.8%	15.0%

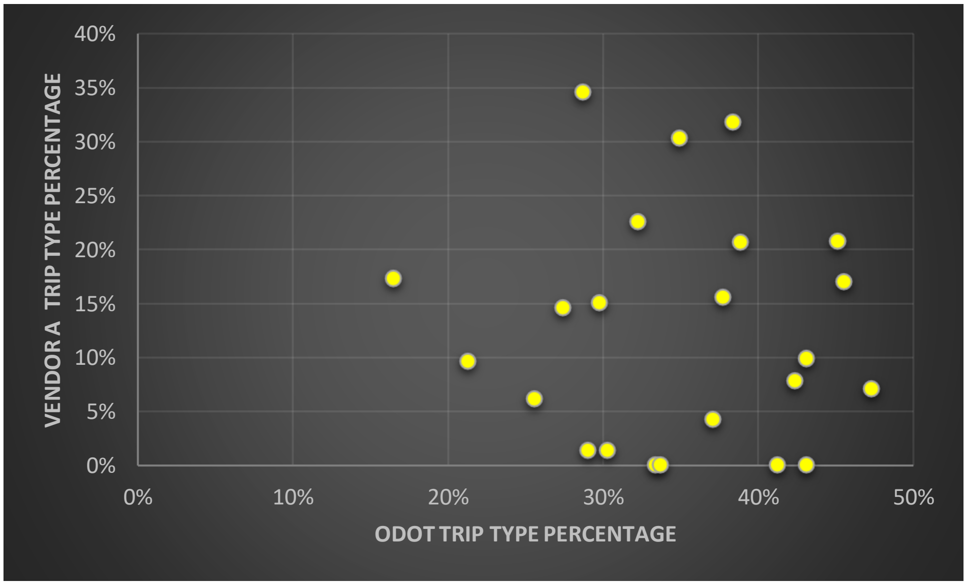


Figure 17: Scatterplot of Vendor A and 2008 ODOT survey commute trip percentages by ingress TAZ

Table 13 extracts the destination TAZs with the largest employment estimates in the study and lists the survey and Vendor A trip type estimates for each. These TAZs have very few residences within their borders and are mostly dominated by one or more major employment centers. Trip types in these TAZs should be heavily weighted towards commute and business types. The table also lists the primary land use or business name for each high employment TAZ. With a handful of exceptions, Vendor A was unable to match the trip type percentages from the survey and logically expected from the land use.

Table 13: Trip type comparison by major employment TAZ destinations

	Business		Personal		Commute		
TAZ	ODOT	Vendor A	ODOT	Vendor A	ODOT	Vendor A	Primary land use
18	6.5%	0.0%	83.9%	90.1%	9.7%	9.9%	Retail
28	13.9%	4.0%	31.6%	49.7%	54.4%	46.2%	Correctional Facility
43	6.3%	4.6%	4.6%	40.2%	89.1%	55.2%	Ford Plant
52	14.3%	0.0%	19.0%	100.0%	66.7%	0.0%	Proctor and Gamble Plant
59	8.4%	3.9%	52.8%	63.8%	38.8%	32.3%	Hospital
66	11.6%	4.6%	72.6%	83.1%	15.8%	12.3%	Retail
67	15.2%	1.9%	63.6%	89.6%	21.2%	8.5%	Retail
99	16.7%	0.0%	13.3%	100.0%	70.0%	0.0%	Manufacturing / Steel plant
100	13.2%	1.3%	12.5%	27.8%	74.3%	70.9%	Husky Fuel Refinery and Depot
103	7.0%	4.9%	14.0%	75.4%	78.9%	19.8%	(JSMC) Tank Manufacturing
134	9.3%	0.0%	44.5%	71.5%	46.2%	28.5%	Hospital
145	15.6%	0.0%	64.6%	98.3%	19.7%	1.7%	Retail
184	12.0%	0.0%	52.0%	100.0%	36.0%	0.0%	Office / Institutional
186	27.6%	0.0%	34.5%	48.9%	37.9%	51.1%	Office / Institutional
287	15.0%	9.1%	26.9%	81.8%	58.1%	9.1%	Industrial / Trucking / Lakeview Farms

Figure 18 shows a scatter plot of the Vendor A and survey commute trip percentages by destination TAZ for the major employment centers. A pattern is apparent in the scatter plot, but the error ranges are large. The results suggest that the trip-type imputation methods used by Vendor A to estimate commute trips may not be able to capture non-standard work hours that may be seen at large manufacturing plants.

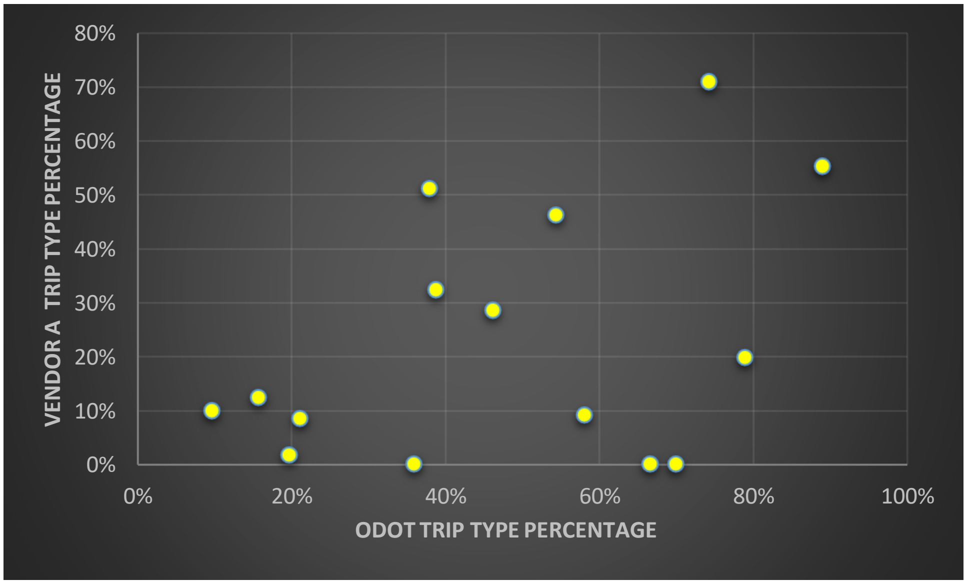


Figure 18: Scatterplot of Vendor A and 2008 ODOT survey commute trip percentages by major employment TAZ destinations

One additional exploration of the Vendor A data was conducted to determine if the spatial resolution of Vendor A's data is impacting the results. Since many of the employment centers may be physically close to the boundary of a TAZ, it is possible that the configuration of cell phone towers may be placing raw data points close to but not exactly on the major center. The Proctor and Gamble manufacturing plant to the east of Lima has over 700 employees and has 12-hour shifts. Figure 19 shows the location of the plant and its distribution center. TAZ boundaries are represented in yellow lines with TAZ #52 highlighted in blue as the plant location. The red circle shows a one-mile radius from the plant (which overlaps with 8 different TAZs). The TAZ IDs are listed labeled in white and the TAZ information is shown in yellow text. The TAZ information contains HW / WH trip percentage and the percentage of trips to/from this TAZ by a designated inbound commuter or outbound commuter (regardless of trip type).

The distribution center TAZ #378 has a low employment value, probably because geocoded employment from this location is placed in TAZ #52 where the manufacturing center is located. Logically, the Vendor A raw data should capture these trips while the survey should show very few. Both Vendor A and the ODOT survey showed 0% of commute trips from this TAZ.

Considering all of the TAZ results within the one-mile radius, Vendor A does not capture the work (HW or WH) external trips for this location. The largest majority of external trip types for this focus area are OO (other-other). It should also be noted that Vendor A is estimating the largest majority of external trips in this focus are traveling to TAZ #46 regardless of trip type.

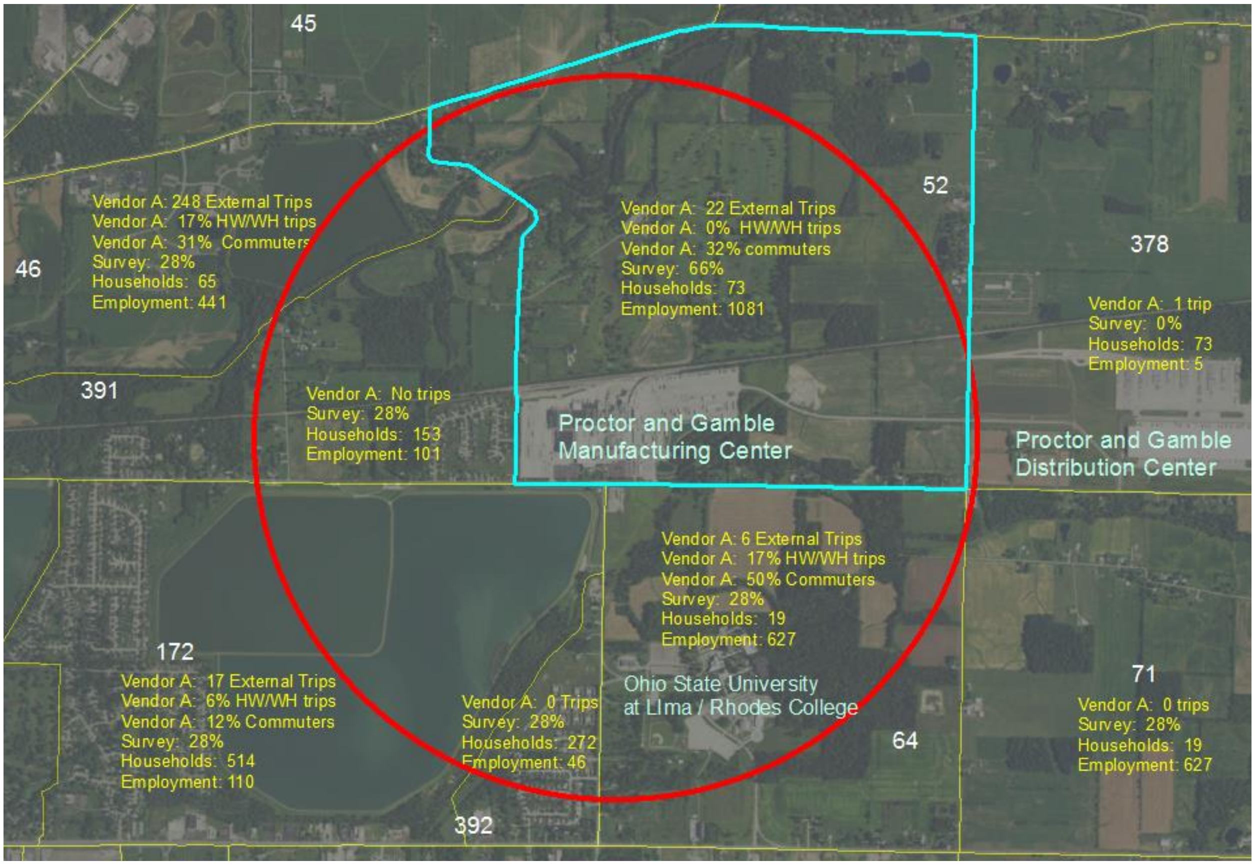


Figure 19: Proctor and Gamble site location

In conclusion, the trip type analysis of the Vendor A dataset shows a modest correlation to the survey and to the major employment centers in Lima. While some success was observed in retail and office/institutional land uses, there is a reason to believe that the trip type imputation algorithms used by Vendor A may not be able to estimate certain commuting patterns and employment centers, possibly due to non-standard working hours. Further, the lack of external trips to the focus area suggests other demographic or location bias exists.

4.2.3 EE flows

Table 14 provides the estimated EE total volume from the different datasets. As can be seen in the table, the data sets have different estimates of total EE volume, especially Vendor B. These values must all be scaled appropriately before analysis or use of their OD estimates. Table 9 provides overall goodness of fit measures for ATD data compared with the appropriate ODOT data. Note in particular the large values for Vendor B data relative to ODOT 2011 Truck data, and compare this to the differences in estimated totals from those datasets in Table 8. R-Square, a measure of correlation, is scale-free and provides a relative goodness of fit assessment. As the R-square values in Table 9 suggest, the relative patterns in the vendor data fit well with the ODOT data for EE flows, with the exception of Vendor A combined with the AADT-weighted assignment system, which shows very poor fit. Again, the fit improvement from applying the shortest path assignment with Vendor A data is dramatic (from 0.05 to 0.93).

Table 14: EE flows - total volume

Data	EE total volume
ODOT 2011 (Total)	27,843
Non-work	14,331
Work	2,130
Truck	11,383
Vendor A: WTD	15,479
Vendor A: SP	2,131
Vendor B: Personal	163,617
Vendor B: Commercial	1,296,875
Vendor C	23,745

Table 15: EE flows - overall goodness of fit measures

ODOT data	ATD	Absolute mean error	Root mean square error	R-Square
ODOT 2011 (total)	Vendor A: WTD	38.9	530.4	0.05
ODOT 2011 (total)	Vendor A: SP	25.8	515.4	0.94
ODOT 2011 Non-work + work	Vendor B: Personal	143.3	2477.3	0.98
ODOT 2011 Truck	Vendor B: Commercial	1241.9	20329.8	0.92
ODOT 2011 Truck	Vendor C	13.4	228.3	0.67

Figure 20, Figure 21 and Figure 22 compare estimated EE flows based on ODOT data versus ATD at the fifteen external stations with the highest proportional flow totals (based on ODOT totals). Figure 20 compares ODOT versus Vendor A estimated EE flows for all vehicles. As Figure 20 suggests, the AADT-weighting method for Vendor A data does a very poor job of reproducing relative pattern of the ODOT totals, while the shortest path assignment method for Vendor A data generates EE flow totals that reproduce well the relative pattern of the ODOT data. This supports the conclusions from the R-square statistics in Table 15.

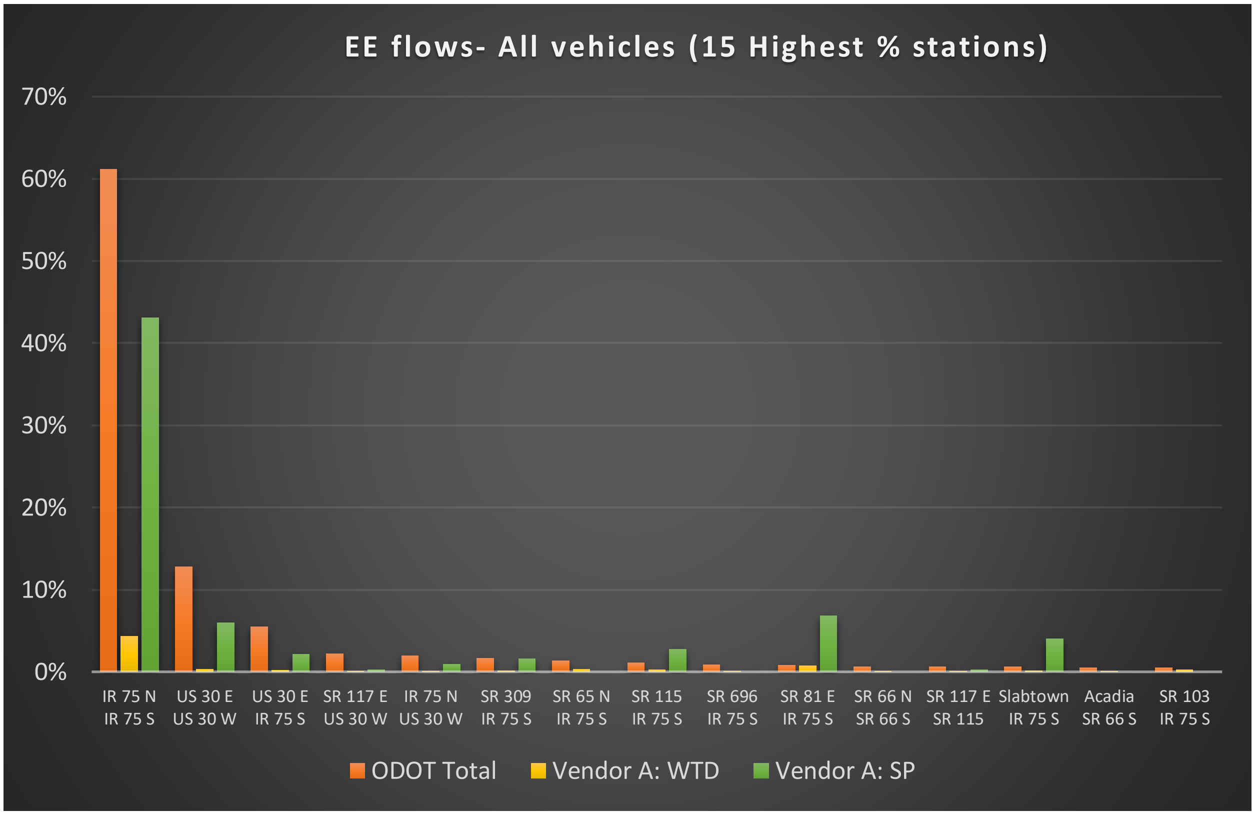


Figure 20: EE flows for all vehicles at high volume external stations: ODOT versus Vendor A

Figure 21 compares ODOT versus Vendor B estimated EE flows for personal vehicles at the fifteen external stations with the highest proportional flow totals (based on ODOT totals). Also supporting the results in Table 15, good correspondence between the ODOT estimates and Vendor B estimates can be seen.

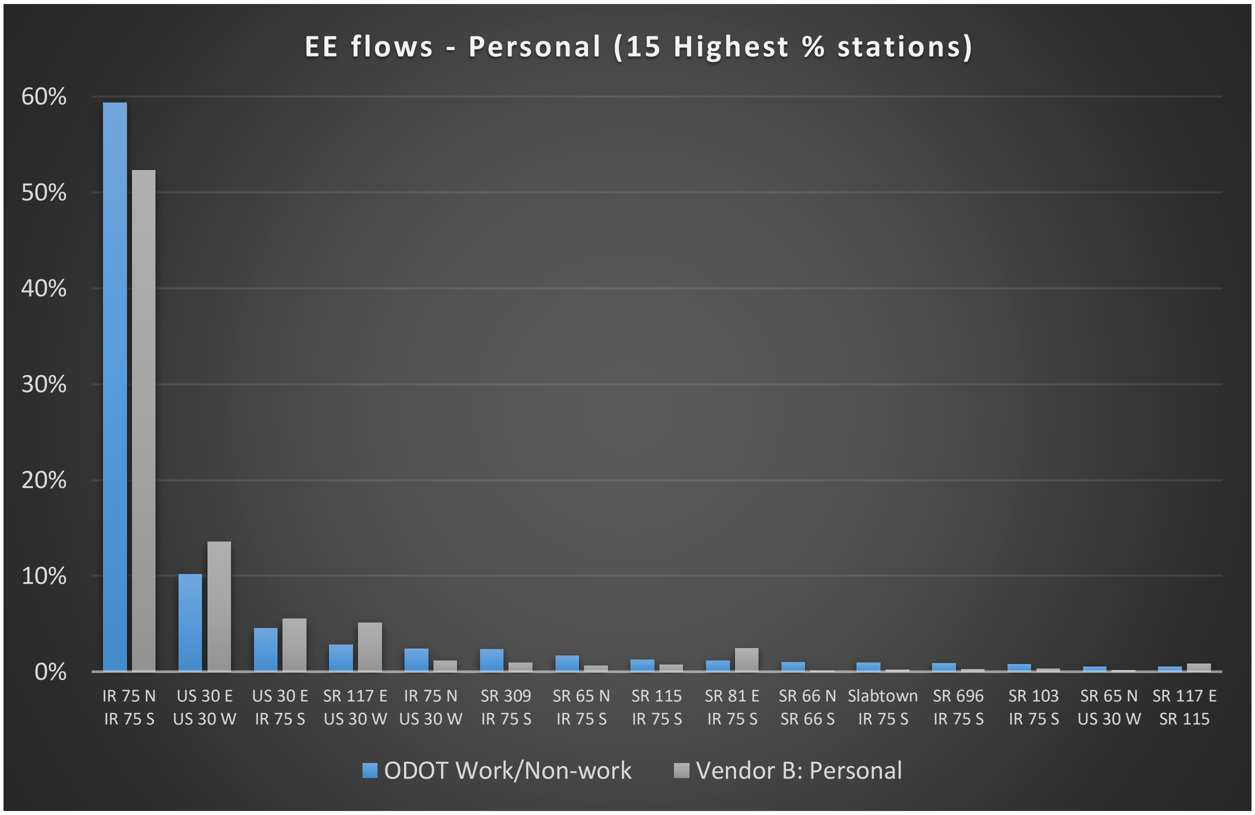


Figure 21: EE flows for personal vehicles at high volume external stations: ODOT versus Vendor B

Finally, Figure 22 compares ODOT versus Vendor B and Vendor C estimated EE flows for commercial vehicles at the fifteen external stations with the highest proportional flow totals (based on ODOT totals). Vendor C did not fit with the ODOT data as well as Vendor B, based on their respective R-square values in Table 15. This may be explained in part by a pattern of higher proportional flows for Vendor C at the IR 75N /US 30 W external station: this is a different relative pattern than is exhibited by the ODOT and Vendor B data, both of which indicate smaller proportional flows across the same sequence of external stations.

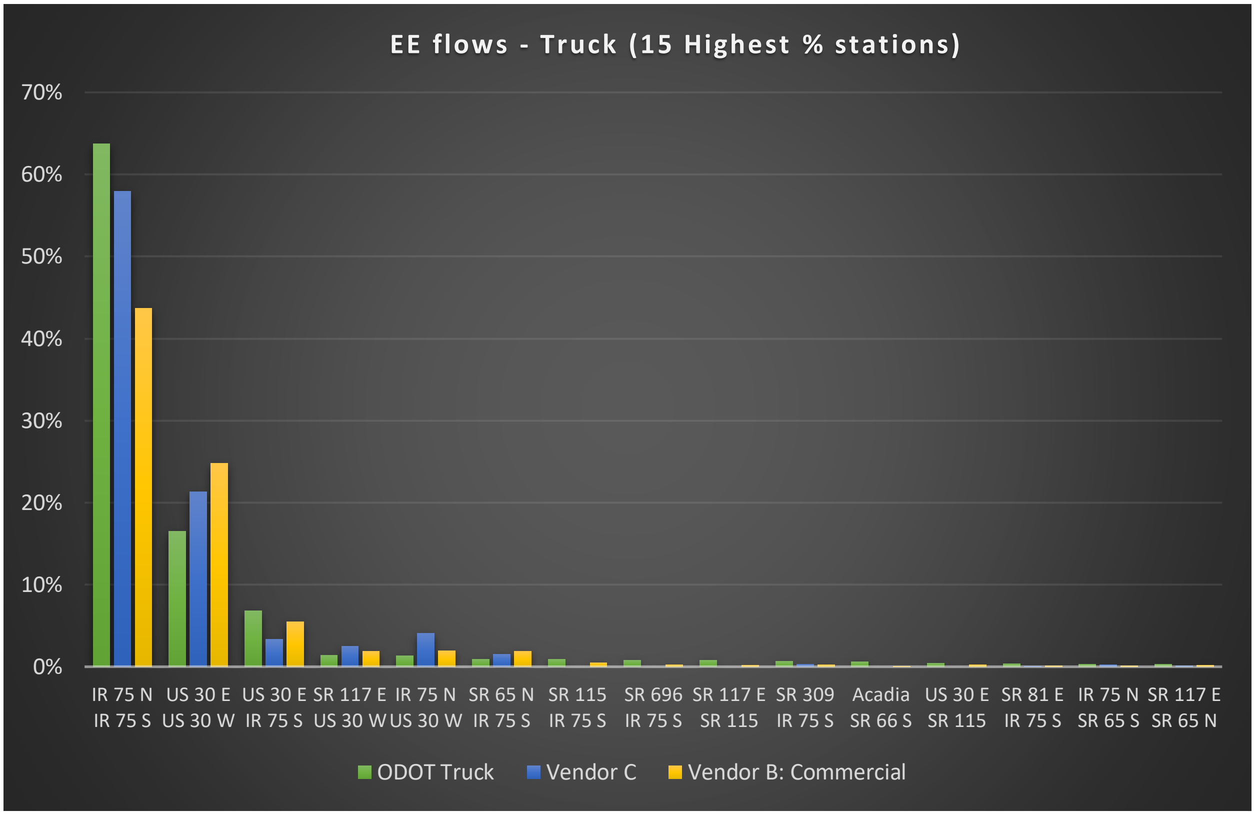


Figure 22: EE flows for commercial vehicles at high volume external stations: ODOT versus Vendor B and Vendor C

The following figures map the EE flows from the datasets Figure 23, Figure 24 and Figure 25 map EE flows for all vehicles based on ODOT, Vendor A: WTD and Vendor A: SP, respectively. As these figures suggest, Vendor A data with the AADT-based weighted assignment method does an extremely poor job of reproducing the similar patterns in the ODOT data (Figure 24), but the same vendor dataset combined with shortest path assignment reproduces the EE flow patterns in the ODOT data well. We can see the dominance of I-75 and (to a lesser degree) US-30 in the EE flows, as expected. These detailed flow maps support the aggregate goodness of fit analysis presented in Table 15.

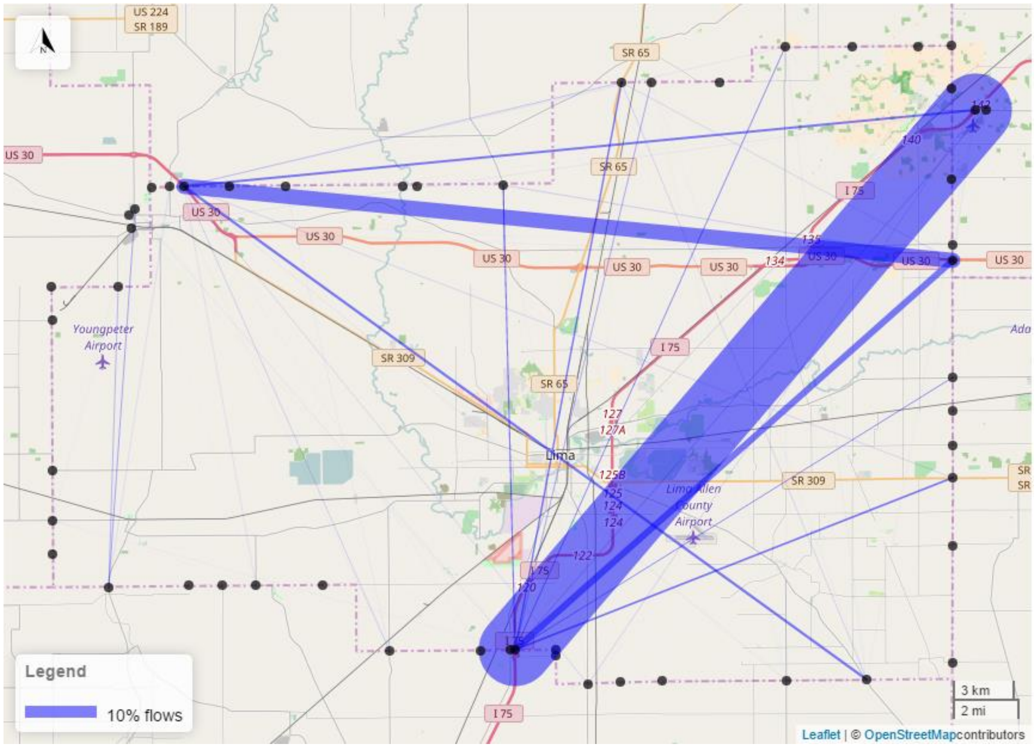


Figure 23: EE flow patterns for all vehicles – ODOT

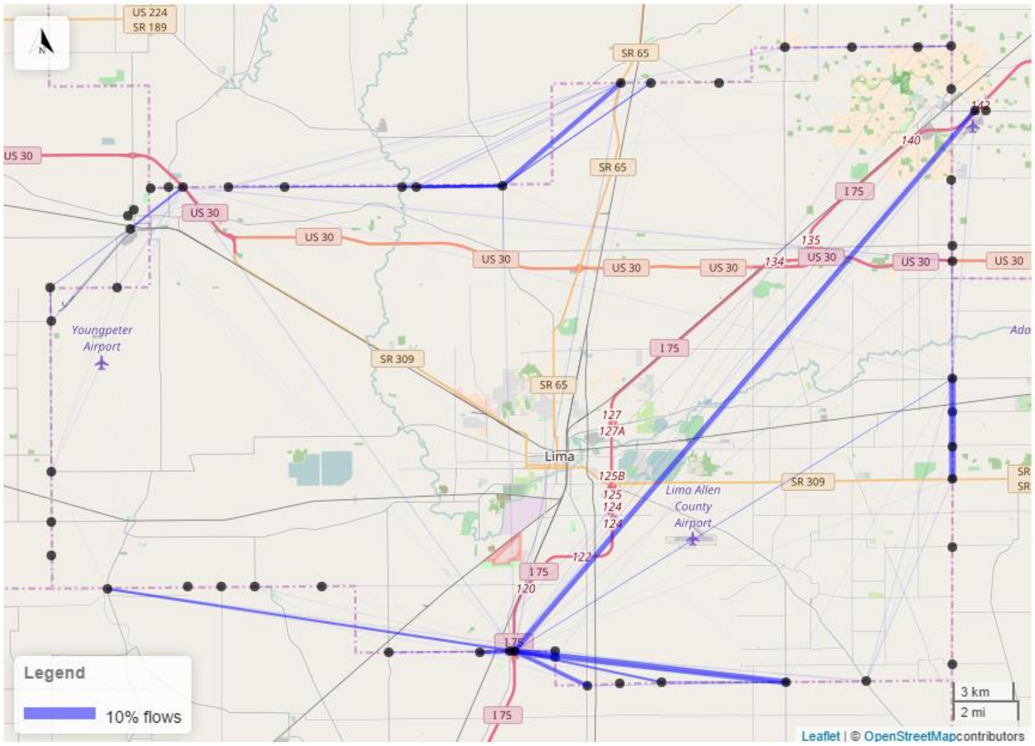


Figure 24: EE flow patterns for all vehicles – Vendor A: WTD

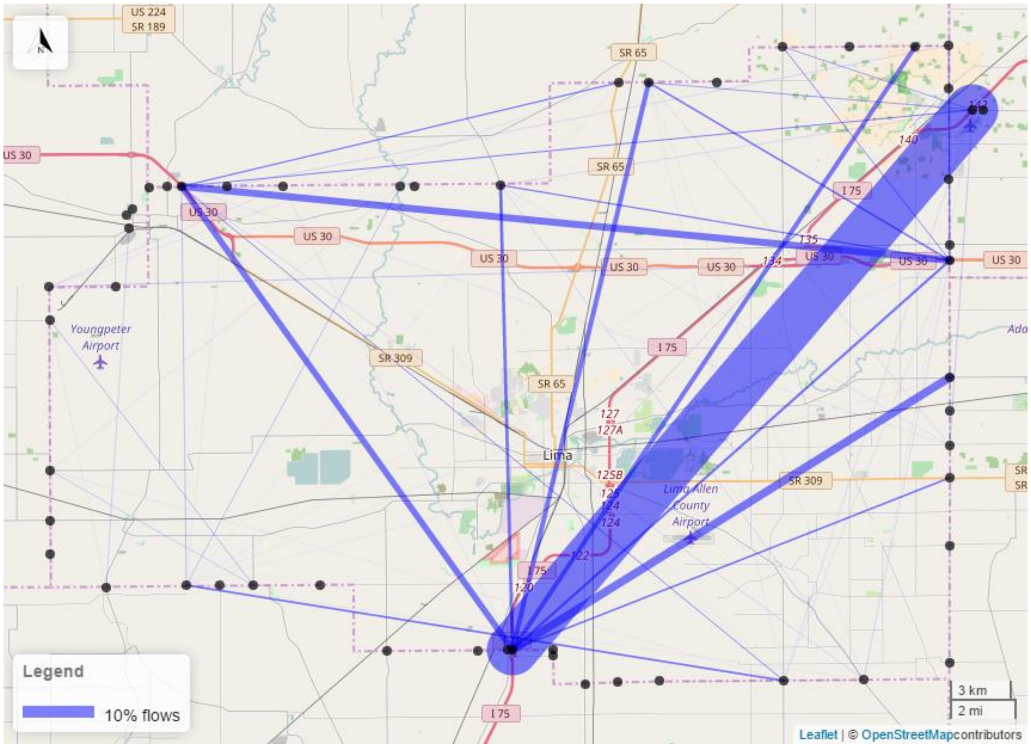


Figure 25: EE flow patterns for all vehicles – Vendor A: SP

Figure 26 and Figure 27 map EE flow patterns for personal vehicles based on ODOT and Vendor B, respectively. These flow patterns appear to match well, supporting the goodness of fit analysis in Table 15).

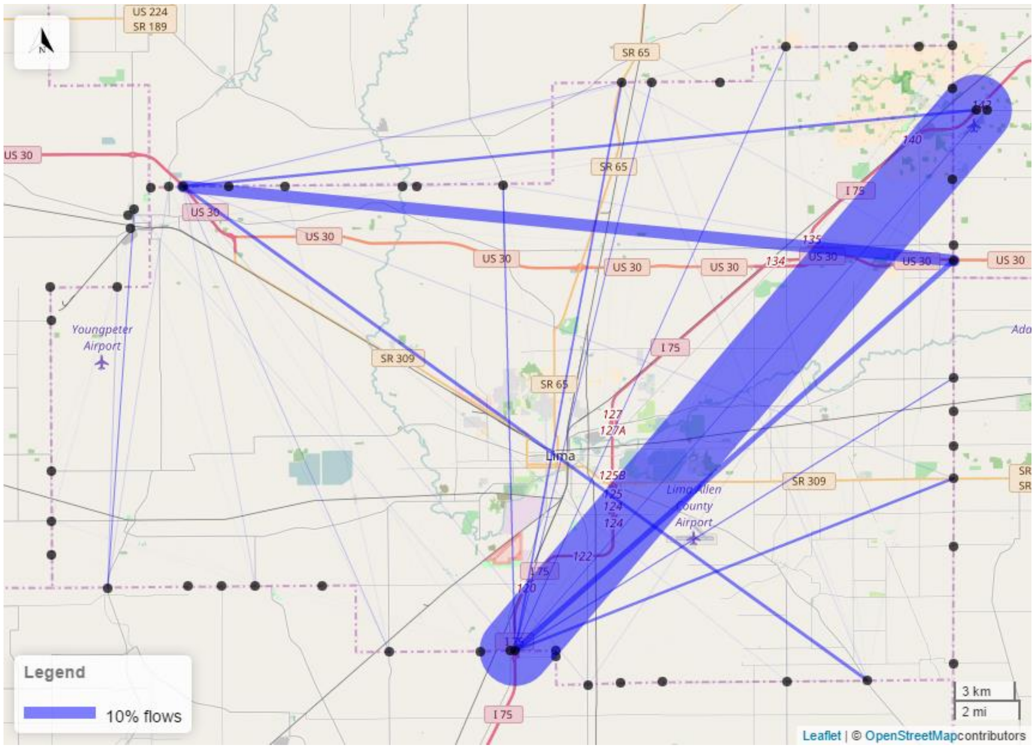


Figure 26: EE flow patterns for personal vehicles – ODOT

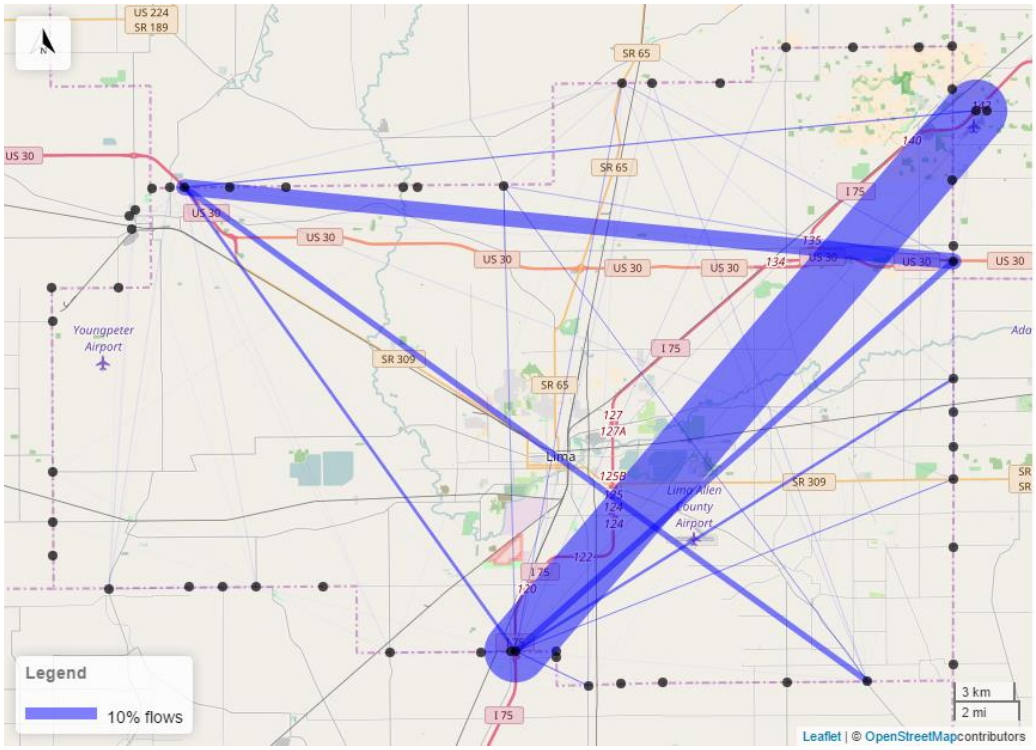


Figure 27: EE flow patterns for personal vehicles - Vendor B

Finally, Figure 28, Figure 29 and Figure 30 map EE flows for trucks based on ODOT, Vendor B and Vendor C data, respectively. All three maps display similar flow patterns dominated by I-75 and US 30, as expected, but with Vendor B and Vendor C showing slightly more dispersed patterns with some flows outside the I-75 and US 30 dominance.

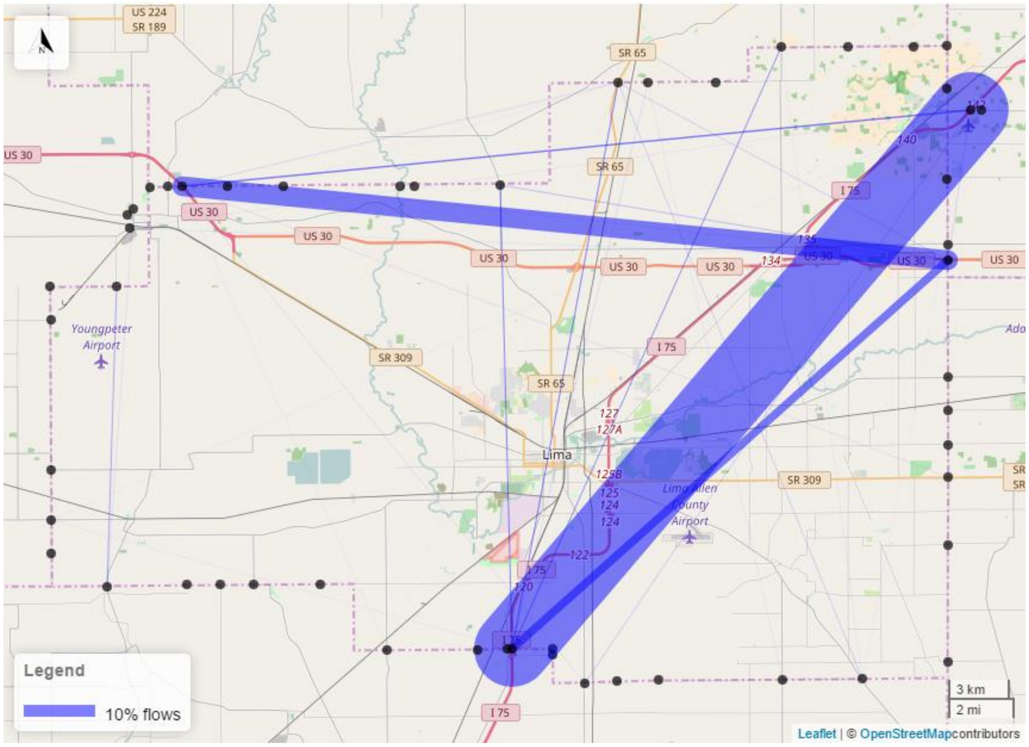


Figure 28: EE flow patterns for trucks – ODOT

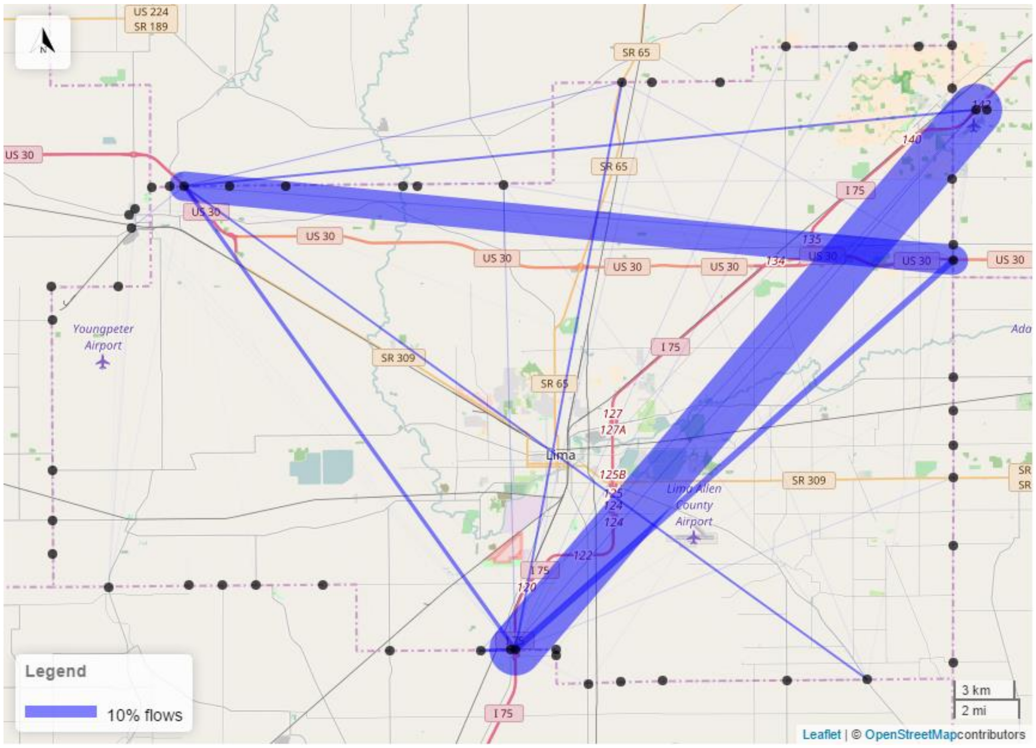


Figure 29: EE flow patterns for trucks - Vendor B

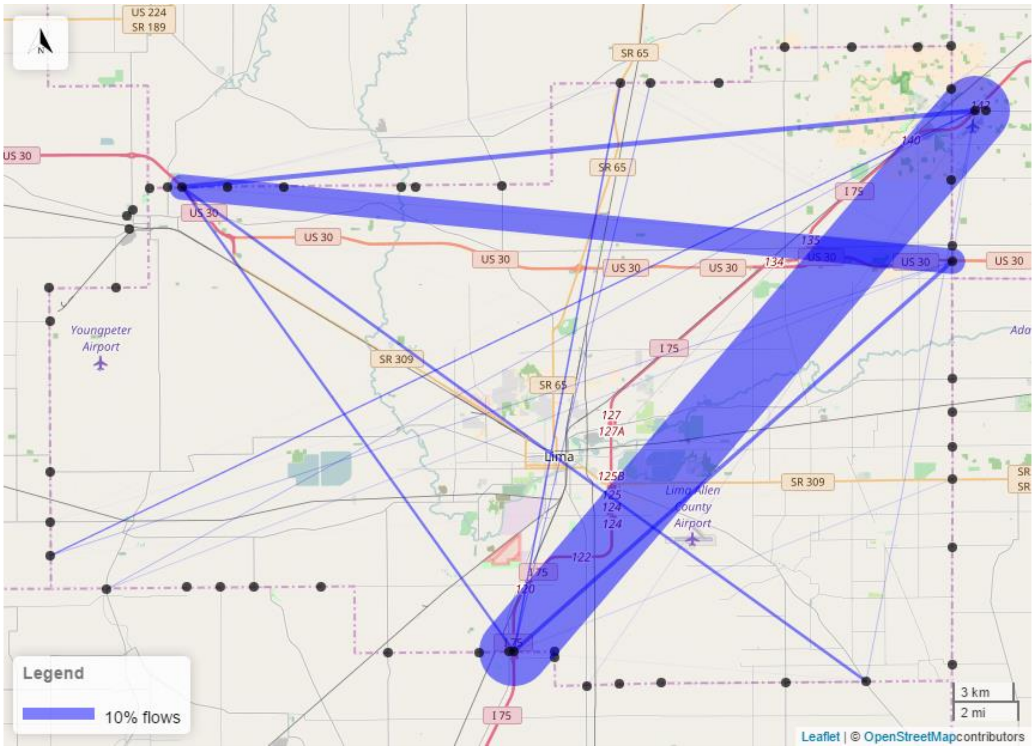


Figure 30: EE flow patterns for trucks - Vendor C

Table 16 provides a comprehensive correlation analysis among all the datasets. This allows for the assessment of consistency not only with ODOT data, but also between the ATDs. For ease of interpretation, table entries are colored red for values less than 0.33, yellow for values between 0.33 and 0.66 inclusive, and green for values greater than 0.66. As Table 16 suggests, most of the datasets appear to tell the same story, suggesting a similar pattern among personal and commercial EE flow patterns regardless of the data source. An exception is Vendor A data with the AADT-based weighting method: this latter dataset is inconsistent with all other datasets.

Table 16: EE flow correlational analysis

EE flow correlations	ODOT 2011 Work	ODOT 2011 Non-Work	ODOT 2011 Truck	ODOT 2011 Total	Vendor A: WTD	Vendor A: SP	Vendor B: Per	Vendor B: Com	Vendor C
ODOT 2011 Work	**1.00**								
ODOT 2011 Non-Work	**0.98**	**1.00**							
ODOT 2011 Truck	**0.97**	**0.99**	**1.00**						
ODOT 2011 Total	**0.98**	**1.00**	**1.00**	**1.00**					
Vendor A: WTD	**0.22**	**0.23**	**0.23**	**0.23**	**1.00**				
Vendor A: SP	**0.95**	**0.95**	**0.95**	**0.95**	**0.23**	**1.00**			
Vendor B: Personal	**0.96**	**0.99**	**0.99**	**0.99**	**0.24**	**0.94**	**1.00**		
Vendor B: Commercial	**0.90**	**0.93**	**0.95**	**0.94**	**0.23**	**0.87**	**0.96**	**1.00**	
Vendor C	**0.95**	**0.97**	**0.99**	**0.98**	**0.22**	**0.93**	**0.99**	**0.98**	**1.00**

Note: Table entries are colored red for values less than 0.33, yellow for values between 0.33 and 0.66 inclusive, and green for values greater than 0.66

4.2.4 EI/IE flows

Table 17 provides the estimates total combined external-internal/internal-external flows based on ODOT data and the unscaled ATD. Table 18 provides the corresponding overall goodness of fit measures. As Table 18 suggests, none of the vendor data fit well with the ODOT data.

Table 17: EI/IE flows - total volume

Data	EI/IE flow total
ODOT 2011 (Total)	98,173
Non-work	54,365
Work	34,660
Truck	9148
Vendor A: WTD	38,477
Vendor A: SP	39,137
Vendor B: Personal	201,473
Vendor B: Commercial	945,254
Vendor C	30,122

Table 18: EI/IE flows - overall goodness of fit measures

Ground-truth	ATD	Absolute mean error	Root mean square error	R Square
ODOT 2011 (total)	Vendor A: WTD	5.0	21.3	0.12
ODOT 2011 (total)	Vendor A: SP	5.3	22.8	0.10
ODOT 2011 Non-work + work	Vendor B: Personal	11.4	97.1	0.11
ODOT 2011 Truck	Vendor B: Commercial	47.5	998.2	0.01
ODOT 2011 Truck	Vendor C	1.8	41.0	0.02

Figure 31, Figure 32 and Figure 33 compare the estimated EI/IE flows at the fifteen highest volume external station (based on ODOT totals). Figure 31 indicates that Vendor A data are poor at reproducing the pattern in the ODOT data for all vehicles. Figure 32 and Figure 33 show similar patterns of poor fit: vendor data on personal vehicles and trucks (respectively) show a qualitatively different pattern than the ODOT data.

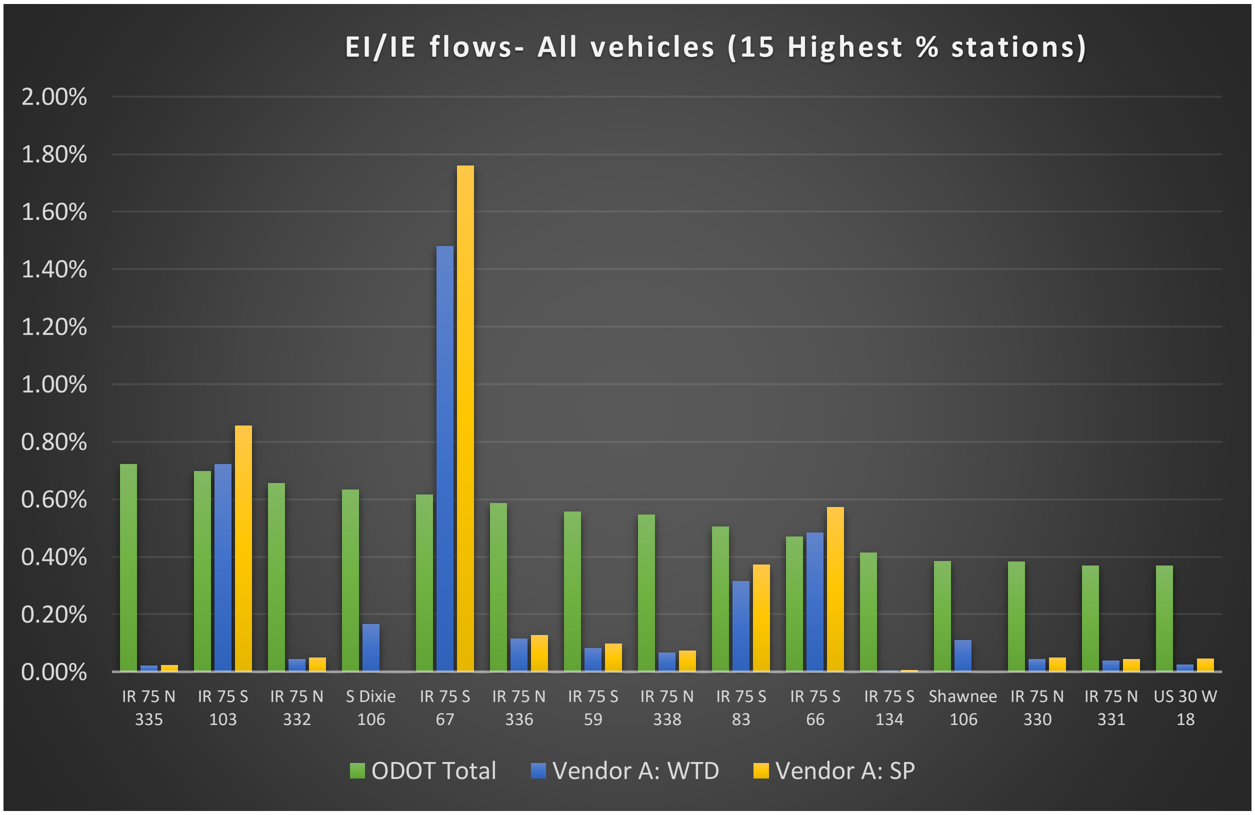


Figure 31: EI/IE flow totals for all vehicles at high volume external stations - ODOT versus Vendor A

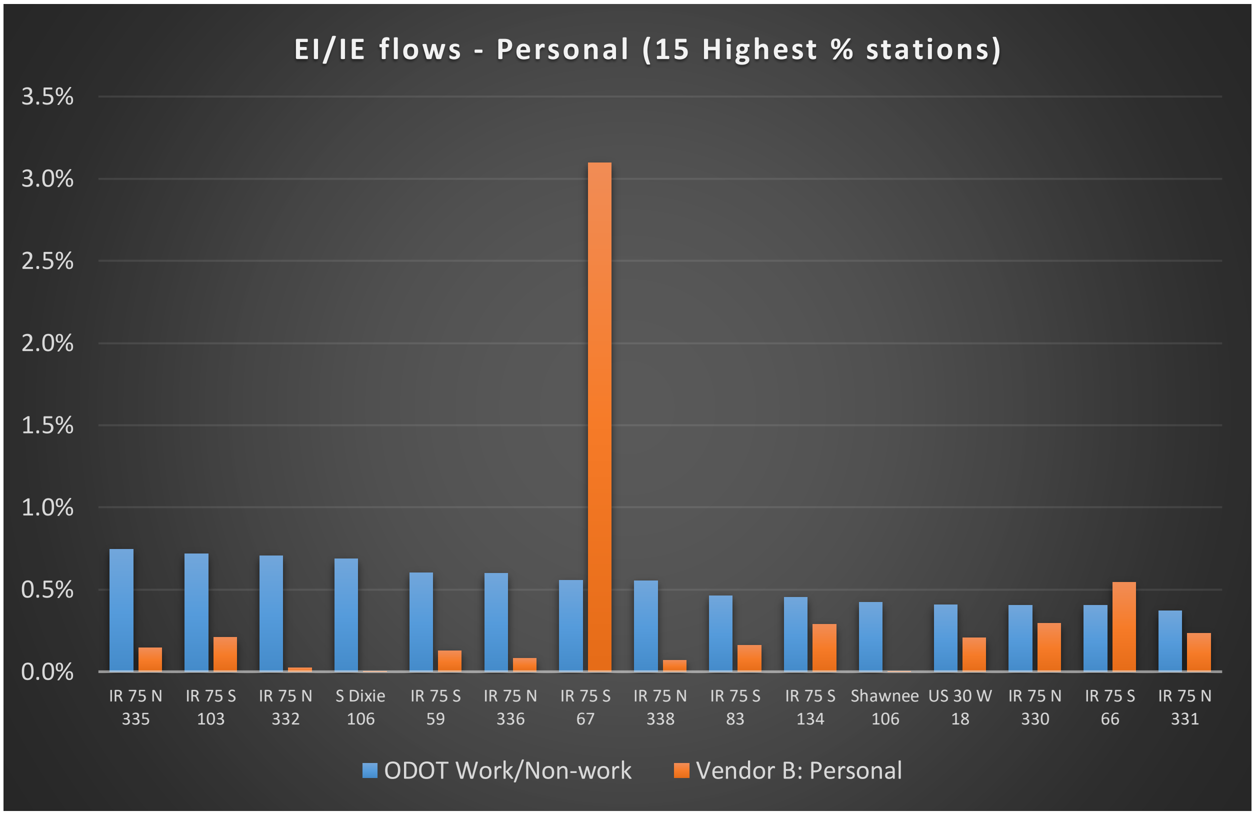


Figure 32: EI/IE flow totals for personal vehicles at high volume external stations - ODOT versus Vendor B

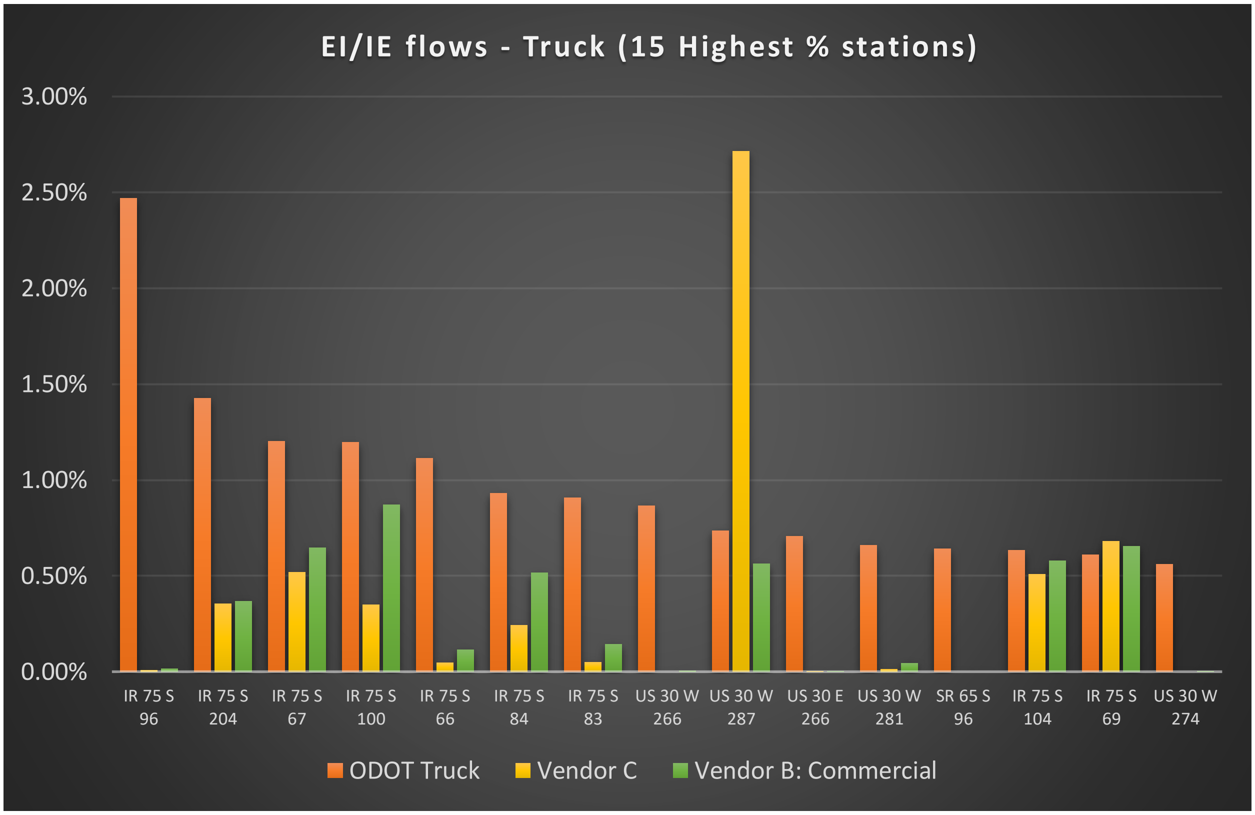


Figure 33: EI/IE flow totals for trucks at high volume external stations - ODOT versus Vendor B and Vendor C

The following figures map the EI/IE flows based on ODOT and vendor data. Figure 34, Figure 35 and Figure 36 map these flows for all vehicles based on ODOT data, Vendor A:WTD and Vendor A:SP, respectively. As Table 18 suggests, the fit between Vendor A data using both assignment methods and the ODOT data is poor. These maps suggest that the ODOT data has a more spatially dispersed EI/IE flow pattern than indicated by Vendor A data.

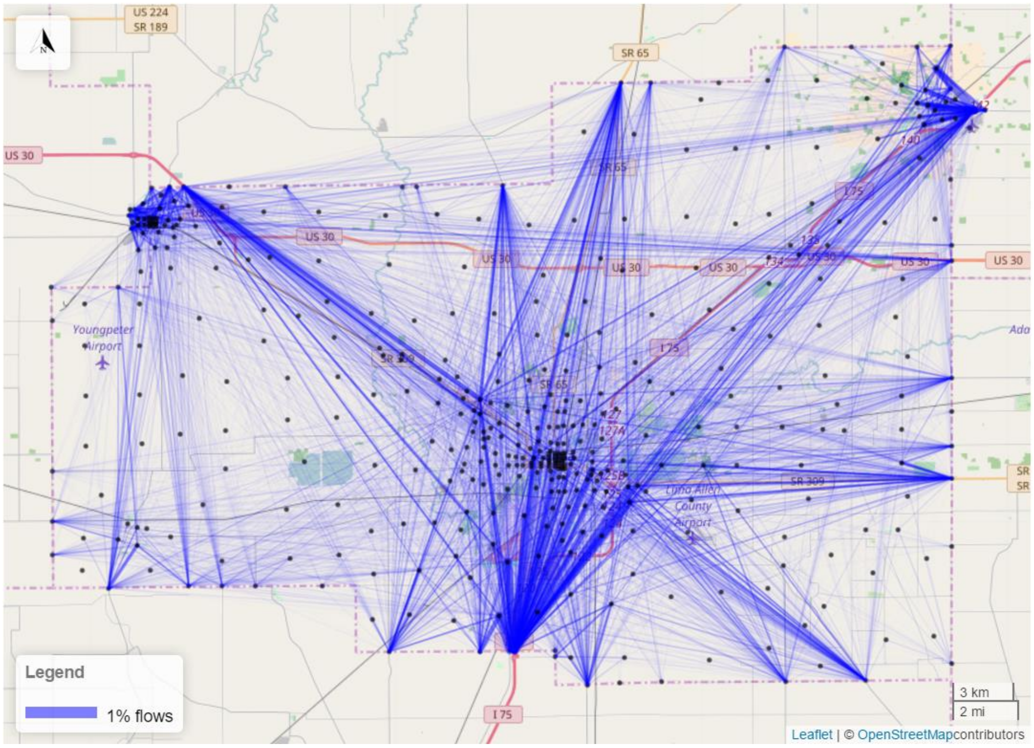


Figure 34: EI/IE flow patterns for all vehicles – ODOT

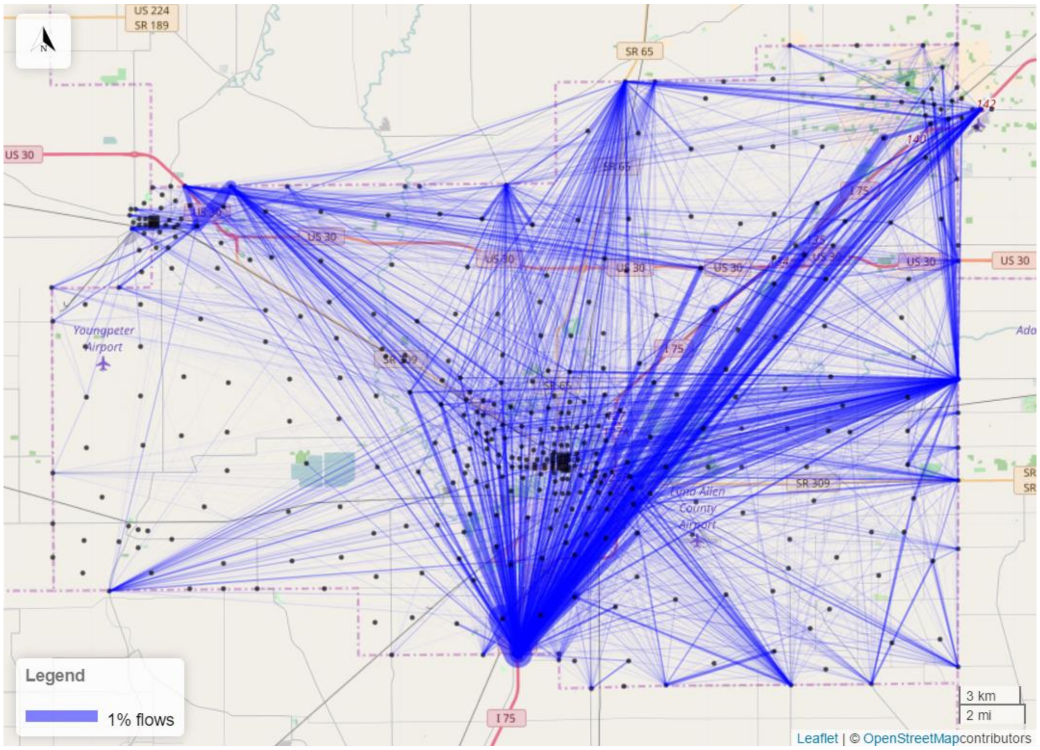


Figure 35: EI/IE flow patterns for all vehicles - Vendor A: WTD

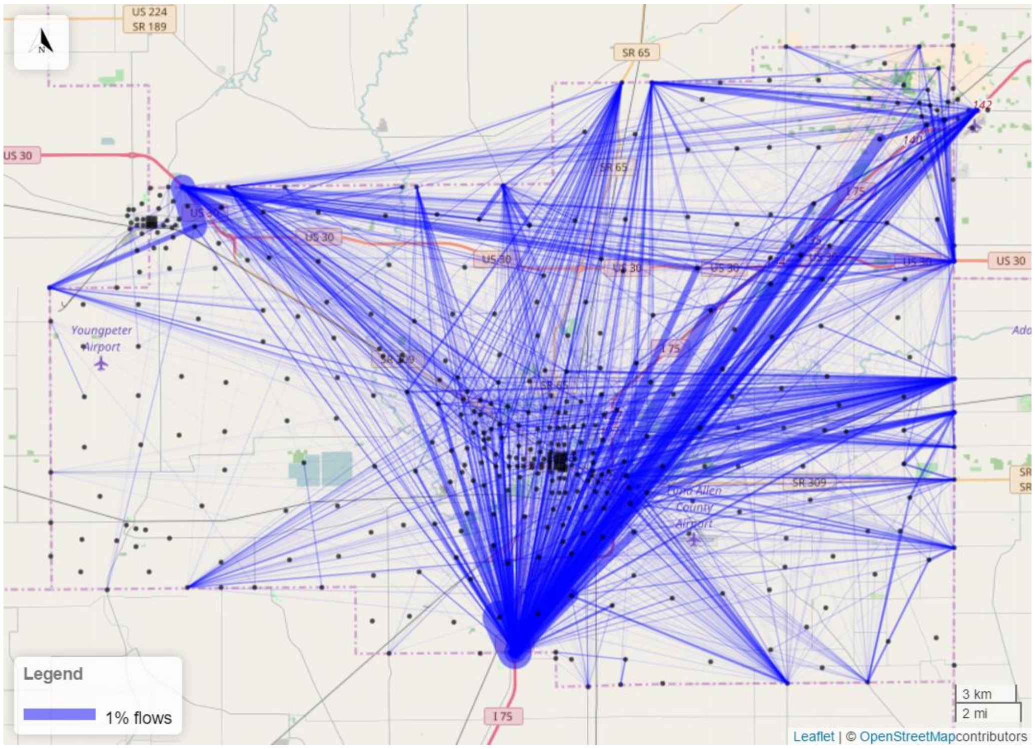


Figure 36: EI/IE flow patterns for all vehicles - Vendor A: SP

Figure 37 and Figure 38 compare EI/IE flow patterns for personal vehicles based on ODOT and Vendor B data, respectively. Recall the poor goodness of fit suggested by the results in Table 18. ODOT data suggest a more spatially dispersed flow pattern (Figure 37) than Vendor B (Figure 38): EI/IE flows based on the latter dataset have stronger flow concentrations along the I-75 and US 30 corridors.

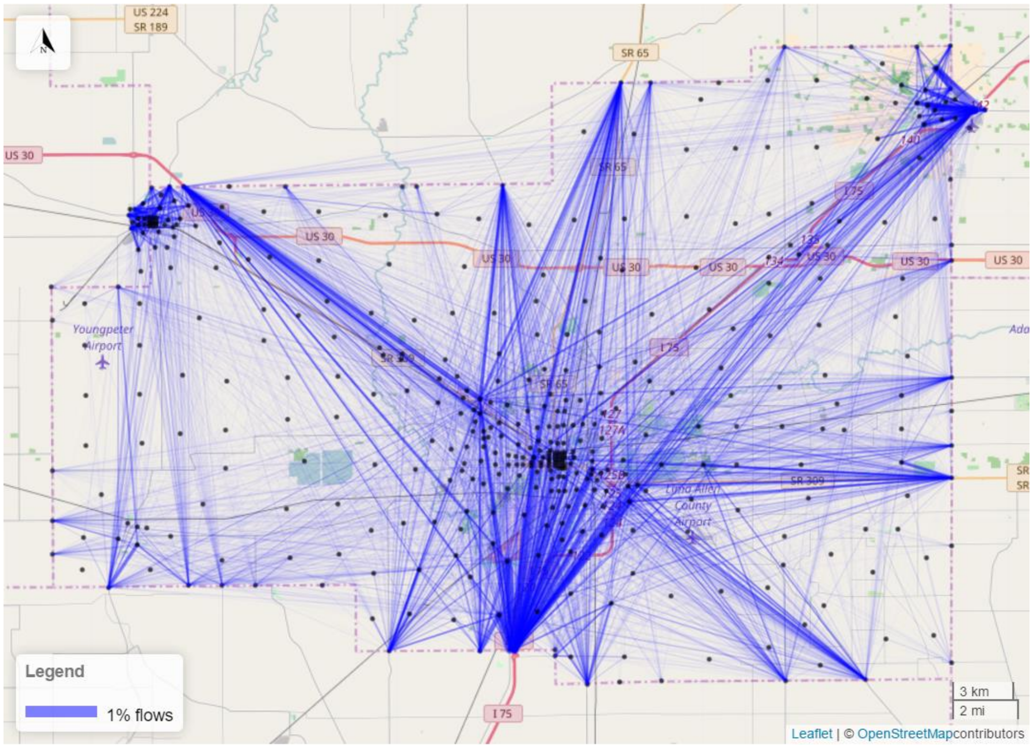


Figure 37: EI/IE flow patterns for personal vehicles – ODOT

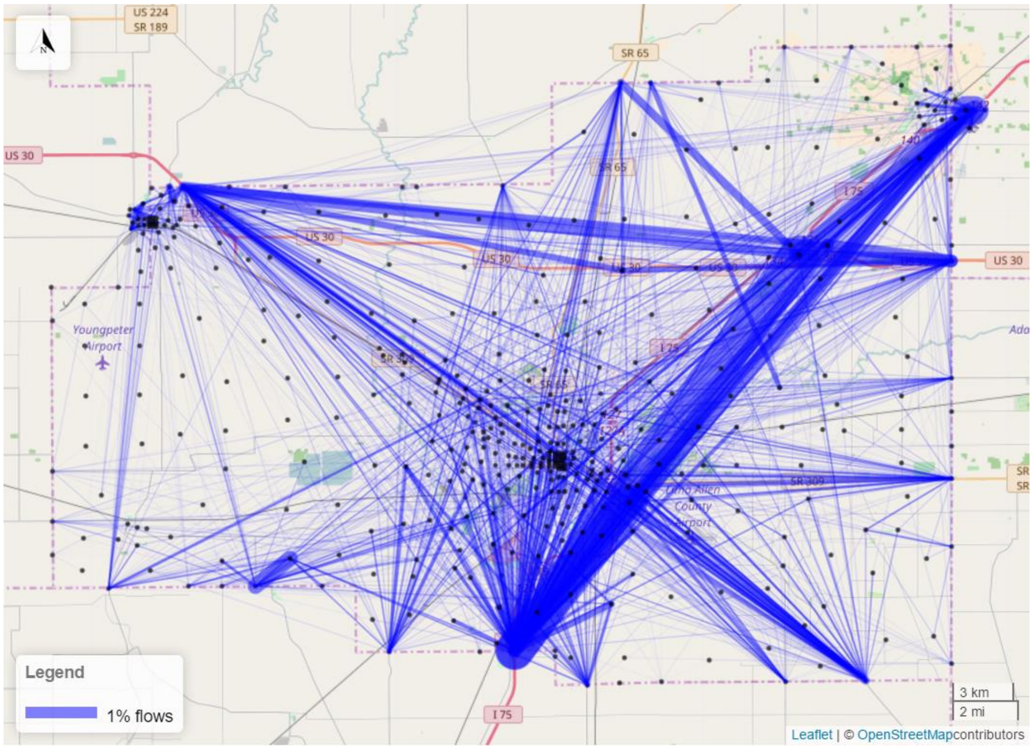


Figure 38: EI/IE flow patterns for personal vehicles - Vendor B

Finally, Figure 39, Figure 40 and Figure 41 map EI/IE flows for trucks based on ODOT, Vendor B and Vendor C data, respectively. Again, recall that goodness of fit statistics in Table 18 suggest poor fit between the vendor and ODOT data. EI/IE truck flows based on ODOT data appear to be more spatially dispersed, albeit with a strong concentration of flows focused on external stations in the northwest corner of the study area (near the town of Delphos). Many of these flows seem to distribute throughout the study area. In contrast, Vendor B data shows a stronger dominance of I-75 and US 30 (Figure 40) while Vendor C data also suggests the dominance of these corridors plus a large number of localized flows between I-75 and US 30 in the northeast portion of the study area as well as where I-75 crosses the Allen County line in the south central portion of the study area.

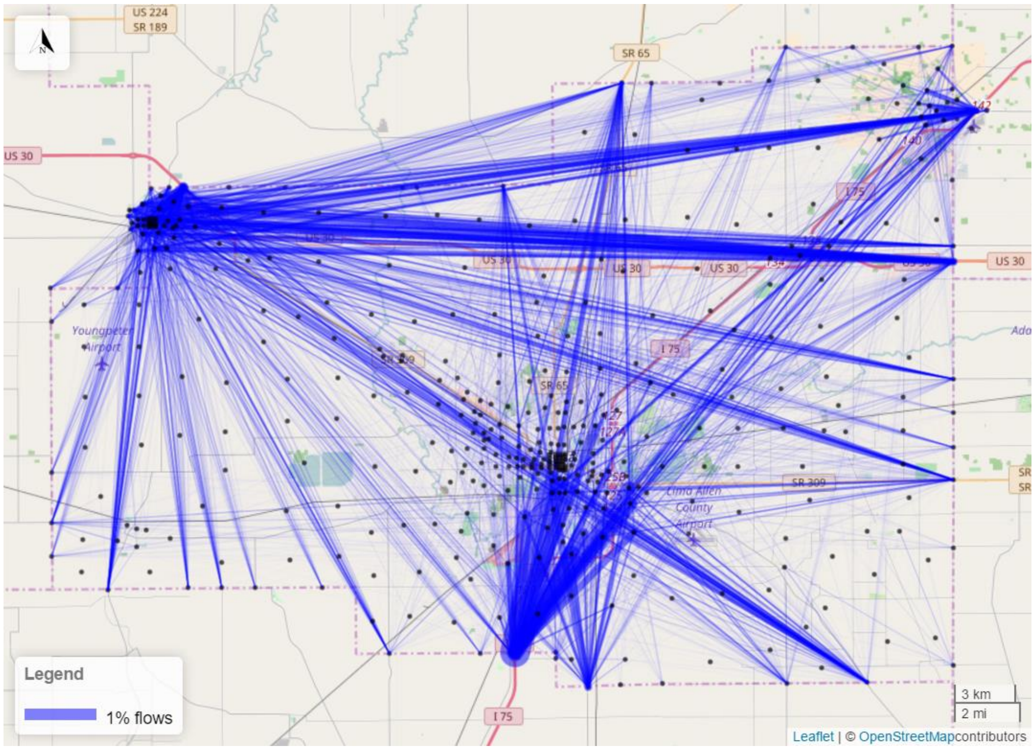


Figure 39: EI/IE flow patterns for trucks – ODOT

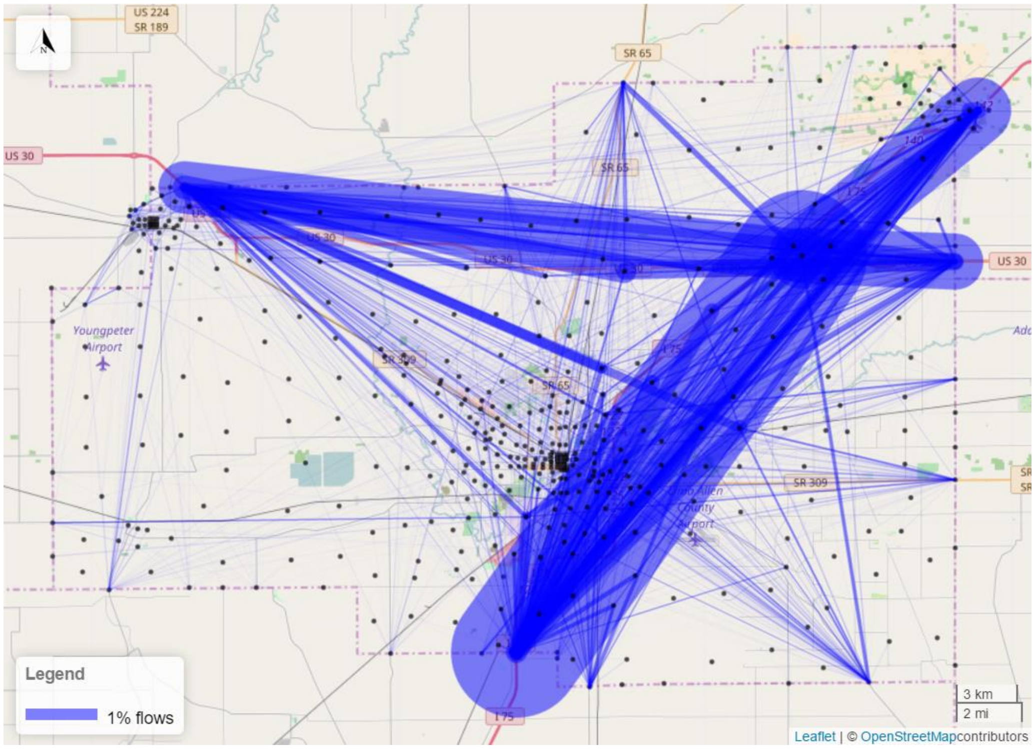


Figure 40: EI/IE flow patterns for trucks - Vendor B

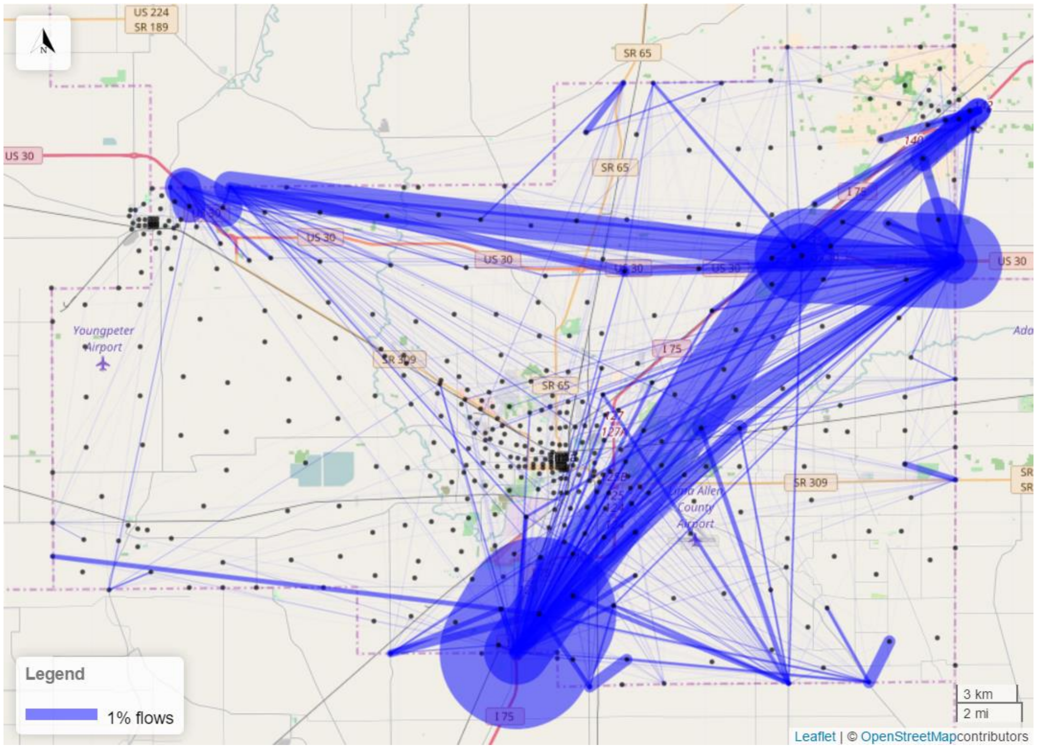


Figure 41: EI/IE flow patterns for trucks - Vendor C

The following set of maps compares spatial patterns of trip count estimates by traffic analysis zone (TAZ). To facilitate comparison, the proportion of total trip counts by TAZ for each dataset was calculated and the data categorized for mapping using quantiles (equal number of observations in each category). Combined maps of the differences between ODOT and ATD trip counts could not be produced due to the different TAZ systems used by the vendors.

Figure 17, Figure 18 and Figure 19 compare estimates of all vehicles using ODOT, Vendor A: WTD and Vendor A: SP data, respectively. As Figure 42 illustrates, the ODOT data suggest a relative spatial concentration of trip counts in the Lima TAZs. In contrast, Figure 43 and Figure 44 suggest a more spatially dispersed pattern of trip counts, with higher proportions in rural TAZs especially in the northeast quadrant of the study area.

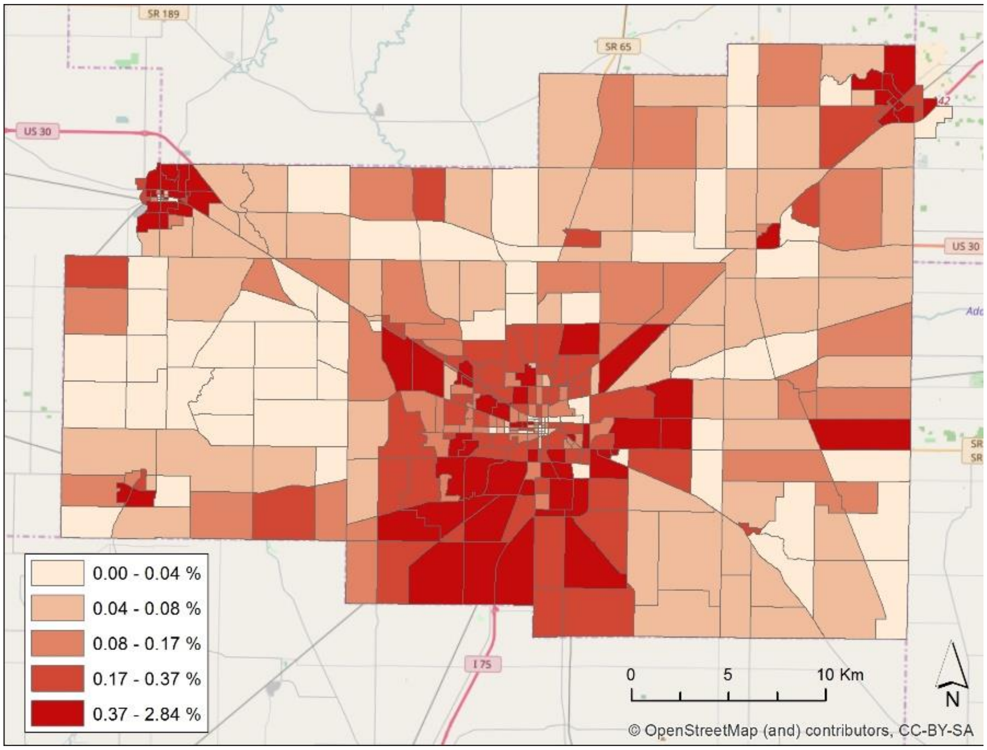
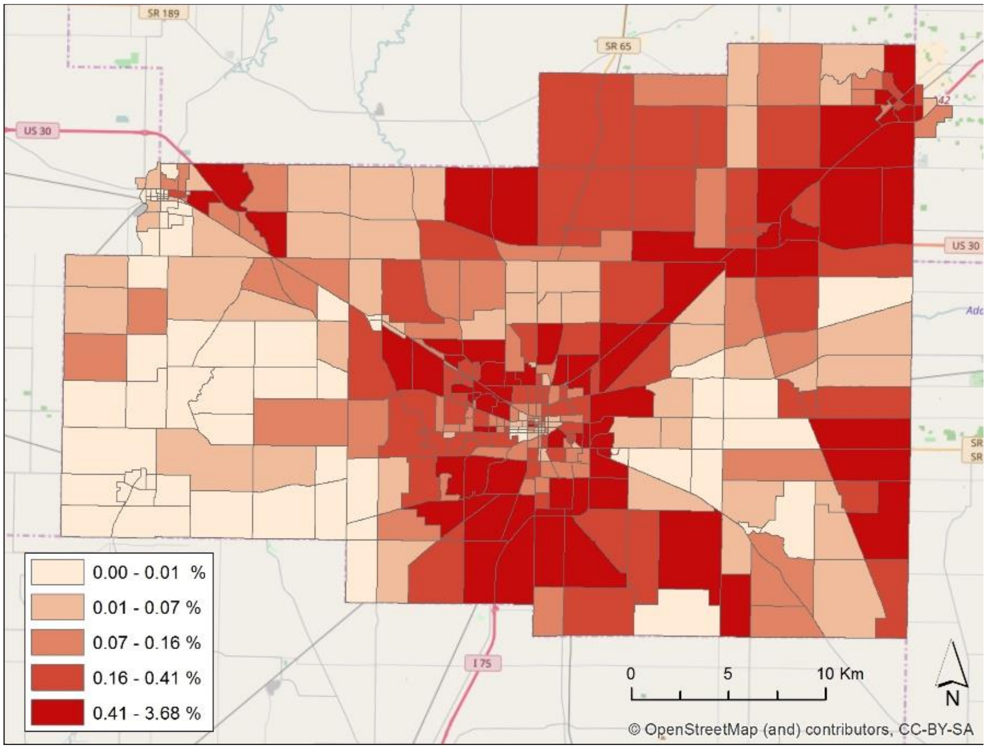
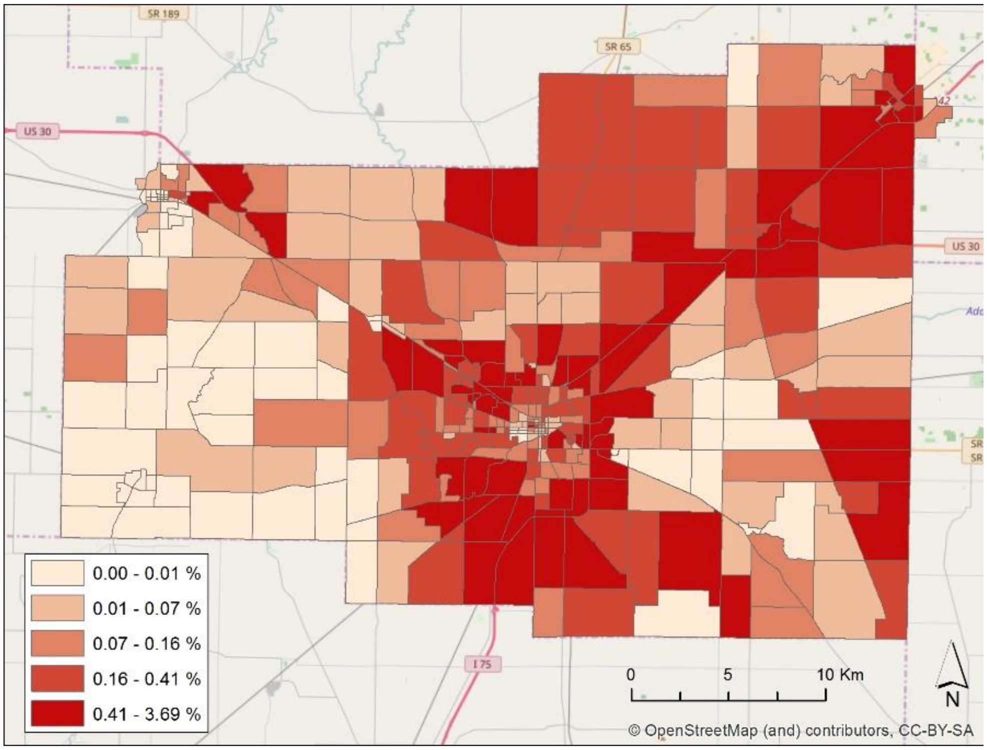


Figure 42: Trip count proportions for all vehicles by TAZs: ODOT



© OpenStreetMap (and) contributors, CC-BY-SA

Figure 43: Trip count proportions for all vehicles by TAZs - Vendor A: WTD



© OpenStreetMap (and) contributors, CC-BY-SA

Figure 44: Trip count proportions for all vehicles by TAZs - Vendor A: SP

Figure 45 and Figure 46 compare proportional trip counts for personal vehicles based on ODOT and Vendor B Personal data. The ODOT personal vehicle trip counts have a similar spatial pattern as the ODOT counts for all vehicles. Trip counts based on Vendor B have a more dispersed, rural pattern, although not as dispersed as the Vendor A data for all vehicles. The Vendor B data also suggests higher trip counts for TAZs in the southern portion of Lima.

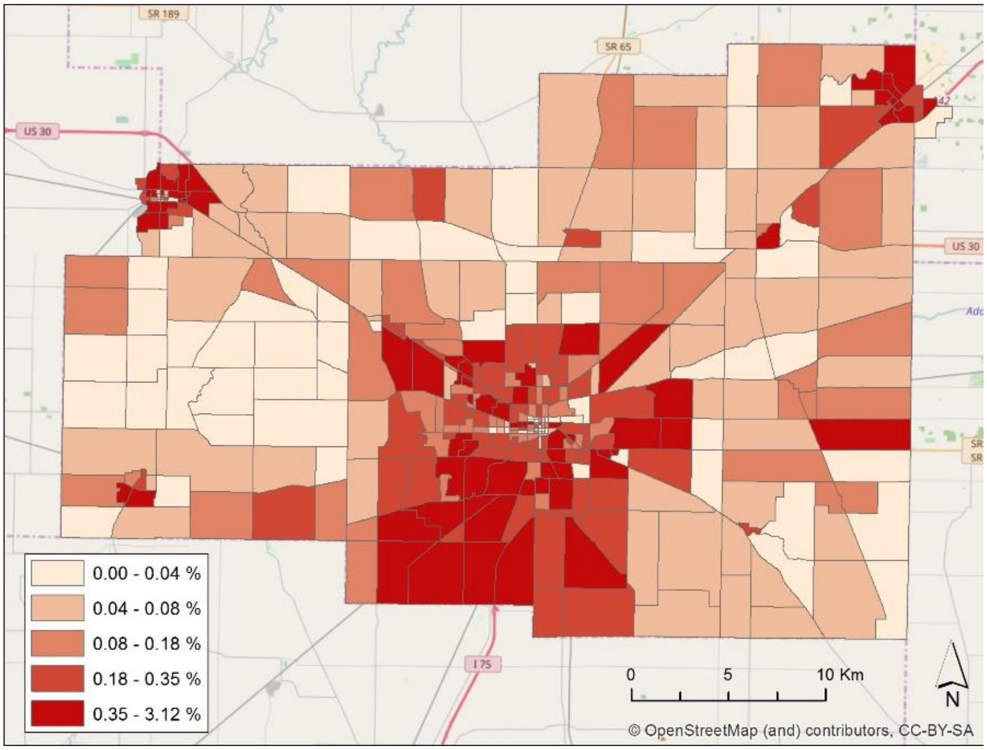


Figure 45: Trip count proportions for personal vehicles by TAZs – ODOT

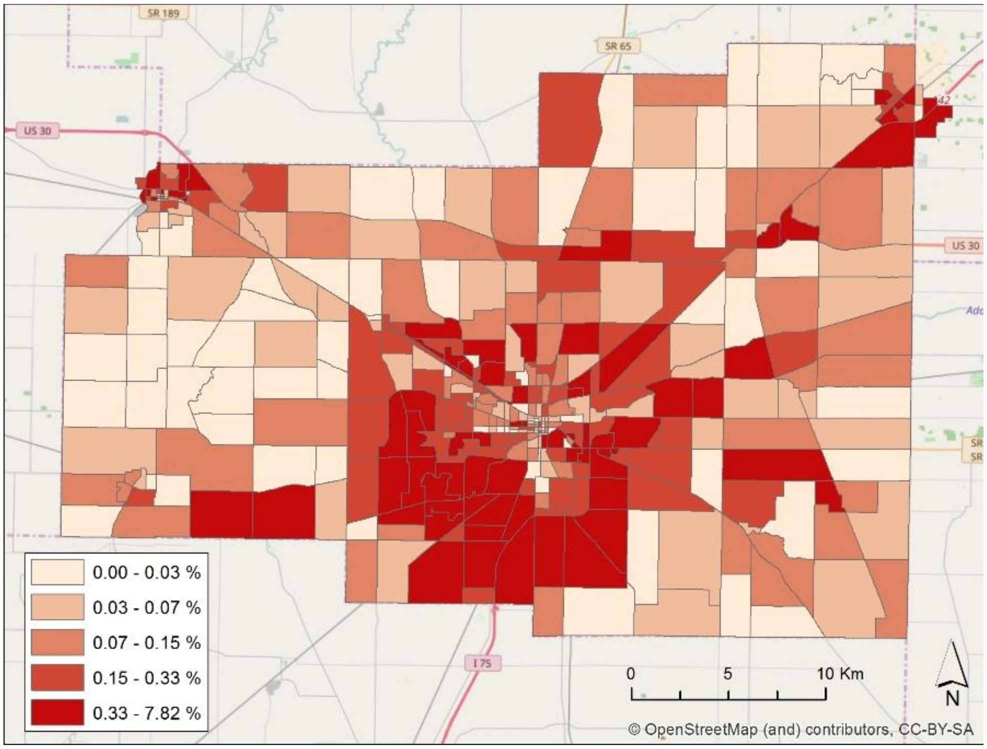


Figure 46: Trip count proportions for personal vehicles by traffic analysis zones - Vendor B

Figure 47, Figure 48 and Figure 49 compare the spatial pattern of proportional trip counts by trucks based on ODOT, Vendor B and Vendor C, respectively. Both Vendor B and Vendor C data indicate higher proportions of trip ends along the in the northern portion of the study area along US-30 and in the northeast quadrant along I-75.

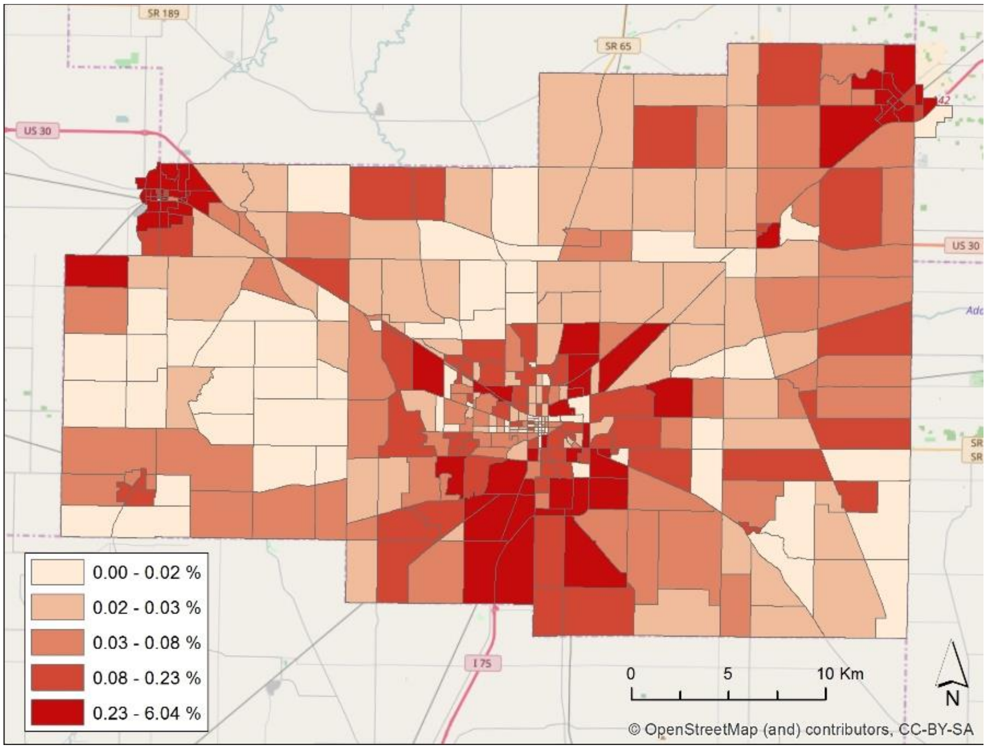


Figure 47: Proportional trip counts for trucks by traffic analysis zones: ODOT

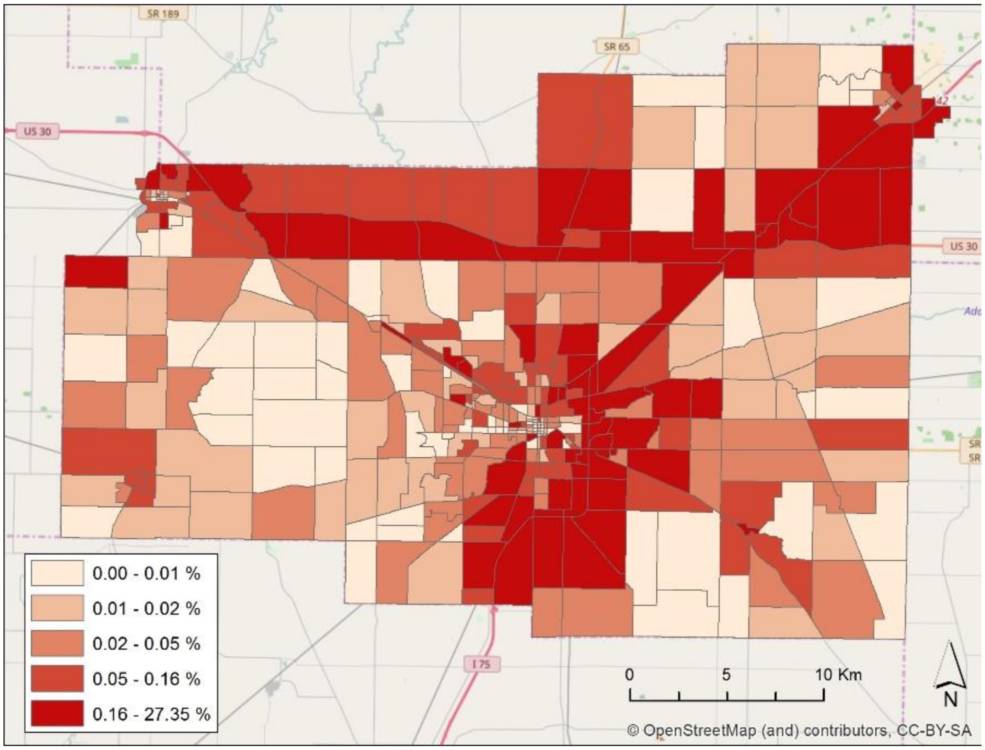


Figure 48: Proportional trip counts for trucks by traffic analysis zone: Vendor B

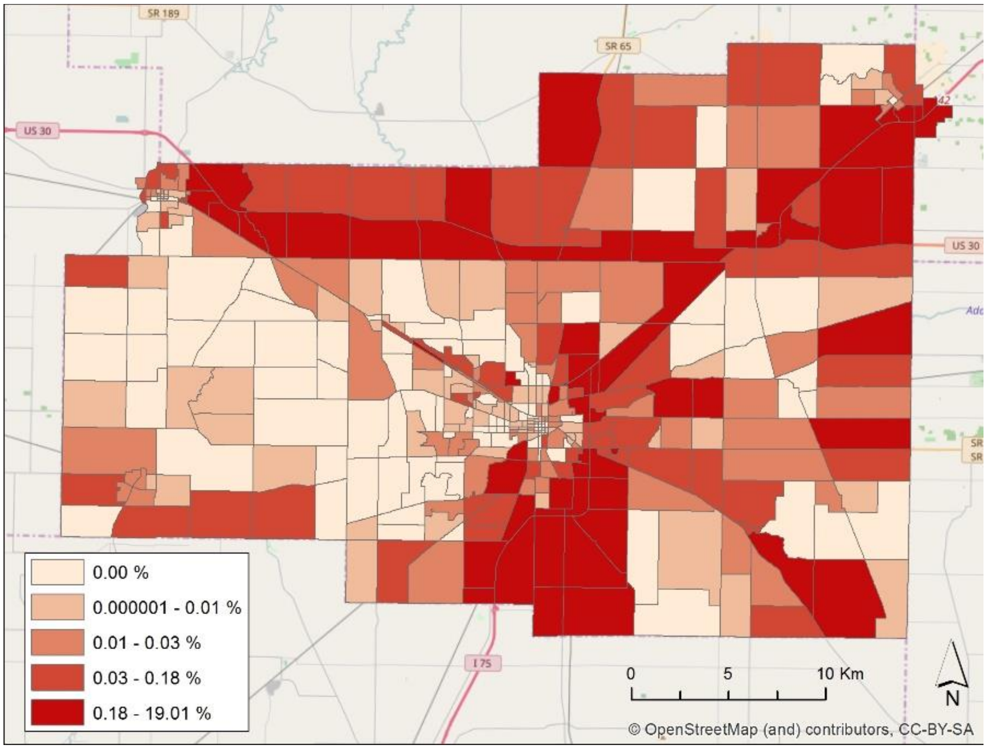


Figure 49: Proportional trip counts for trucks by traffic analysis zones: Vendor C

Table **19** and Table 20 provide comprehensive correlational analyses for EI and IE flows, respectively. As these tables suggest, all of the datasets appear to be telling different stories: the ATD is not only inconsistent with ODOT data, but ATD are also inconsistent with each other.

Table 19: EI flow correlational analysis

EI flow correlations	ODOT 2011 Work	ODOT 2011 Non-Work	ODOT 2011 Truck	ODOT 2011 Total	Vendor A: WTD	Vendor A: SP	Vendor B: Per	Vendor B: Com	Vendor C
ODOT 2011 Work	**1.00**								
ODOT 2011 Non-Work	**0.60**	**1.00**							
ODOT 2011 Truck	**0.36**	**0.36**	**1.00**						
ODOT 2011 Total	**0.87**	**0.91**	**0.49**	**1.00**					
Vendor A: WTD	**0.22**	**0.27**	**0.28**	**0.29**	**1.00**				
Vendor A: SP	**0.21**	**0.25**	**0.28**	**0.27**	**0.89**	**1.00**			
Vendor B: Personal	**0.18**	**0.42**	**0.30**	**0.36**	**0.41**	**0.41**	**1.00**		
Vendor B: Commercial	**0.11**	**0.13**	**0.15**	**0.15**	**0.31**	**0.32**	**0.61**	**1.00**	
Vendor C	**0.04**	**0.09**	**0.10**	**0.08**	**0.20**	**0.20**	**0.22**	**0.31**	**1.00**

Note: Table entries are colored red for values less than 0.33, yellow for values between 0.33 and 0.66 inclusive, and green for values greater than 0.66

Table 20: IE flow correlational analysis

IE flow correlations	ODOT 2011 Work	ODOT 2011 Non-Work	ODOT 2011 Truck	ODOT 2011 Total	Vendor A: WTD	Vendor A: SP	Vendor B: Pers	Vendor B: Com	Vendor C
ODOT 2011 Work	**1.00**								
ODOT 2011 Non-Work	**0.89**	**1.00**							
ODOT 2011 Truck	**0.31**	**0.22**	**1.00**						
ODOT 2011 Total	**0.95**	**0.97**	**0.41**	**1.00**					
Vendor A: WTD	**0.30**	**0.20**	**0.29**	**0.28**	**1.00**				
Vendor A: SP	**0.27**	**0.18**	**0.30**	**0.25**	**0.87**	**1.00**			
Vendor B: Personal	**0.17**	**0.15**	**0.25**	**0.20**	**0.33**	**0.33**	**1.00**		
Vendor B: Commercial	**0.04**	**0.03**	**0.09**	**0.05**	**0.20**	**0.20**	**0.69**	**1.00**	
Vendor C	**0.03**	**0.02**	**0.06**	**0.04**	**0.15**	**0.16**	**0.46**	**0.69**	**1.00**

Note: Table entries are colored red for values less than 0.33, yellow for values between 0.33 and 0.66 inclusive, and green for values greater than 0.66

A similar correlation analysis was performed comparing EI and IE flows with the socioeconomic variables of population and number of workers, using data for TAZs provided by ODOT. However, this analysis required reconciliation of differing TAZ geographies. Specifically, while ODOT and Vendor A data are based on a TAZ system that consists of 378 zones, Vendor B and C data are aggregated to another TAZ system of 395 zones. To reconcile this discrepancy, two zoning maps were overlaid to show that some boundaries are added and edited in the 'TAZ395 system'; see Figure 50. To reconcile this, a TAZ system that consists of 371 zones was created. This was accomplished by aggregating the segmented TAZs to the original parent zone. The number of TAZs are less than 378 because it was necessary to aggregate TAZs in the 'TAZ378 system' to the new bigger zones where the TAZ boundaries was changed.

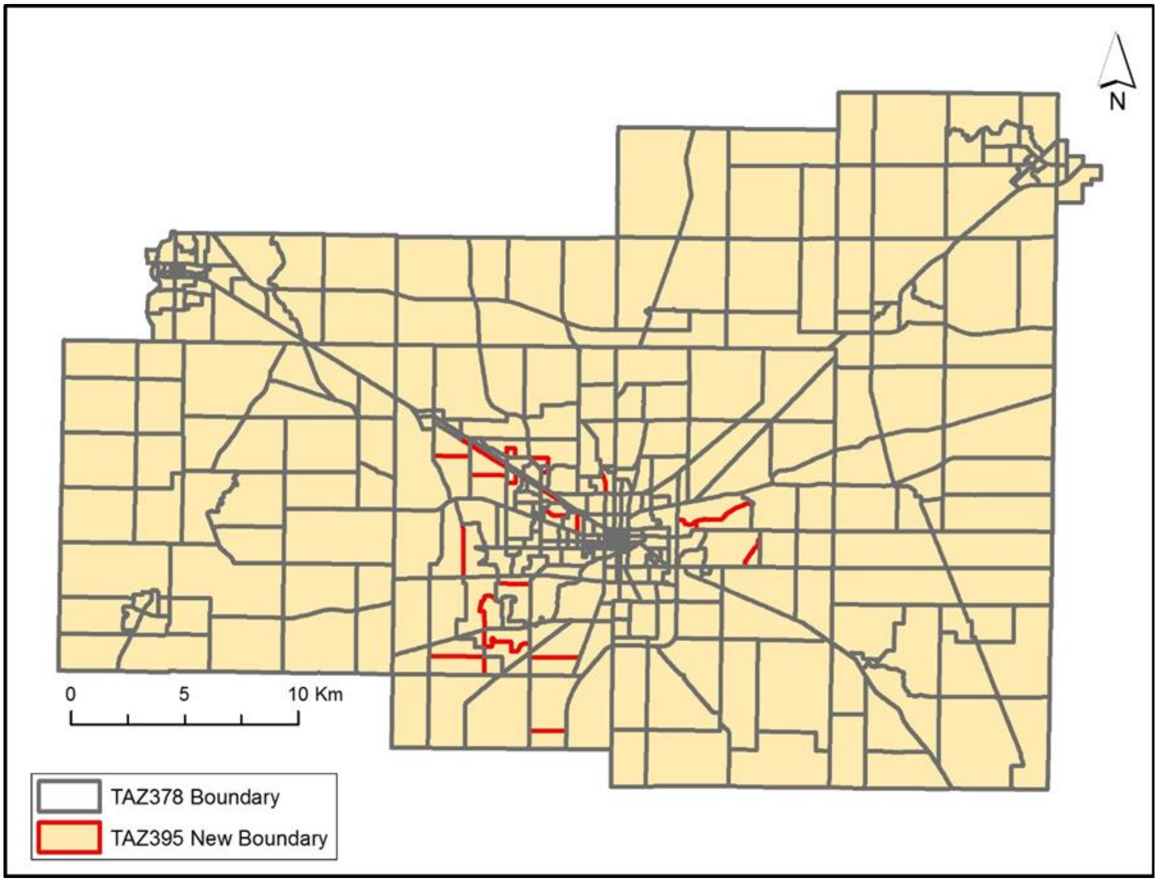


Figure 50: Different TAZ geographies

Table 21 and Table 22 provide the results from the correlational analysis of EI and IE flows (respectively) with the socioeconomic variables of population and number of workers. EI flows have weak correlation with EI flows across all datasets (see Table 21). In contrast, IE flows derived from ODOT 2011 Work, 2011 Non-work and 2011 Total datasets have moderate correlation with population and the number of workers (see Table 22). However, IE flows from the 2011 ODOT Truck and the vendor datasets have weak correlation with population and the number of workers.

Table 21: EI flow and socioeconomic variables correlational analysis

	ODOT 2011 Work	ODOT 2011 Non-Work	ODOT 2011 Truck	ODOT 2011 Total	Vendor A: WTD	Vendor A: SP	Vendor B: Per	Vendor B: Com	Vendor C

Population	0.001	0.000	0.001	0.000	0.062	0.062	0.009	0.002	0.001
Workers	0.001	0.000	0.001	0.000	0.062	0.062	0.016	0.002	0.000
Note: Table entries are colored red for values less than 0.33, yellow for values between 0.33 and 0.66 inclusive, and green for values greater than 0.66									

Table 22: IE flow and socioeconomic variables correlational analysis

	ODOT 2011 Work	ODOT 2011 Non-Work	ODOT 2011 Truck	ODOT 2011 Total	Vendor A: WTD	Vendor A: SP	Vendor B: Per	Vendor B: Com	Vendor C
Population	0.507	0.410	0.001	0.404	0.071	0.071	0.011	0.002	0.002
Workers	0.533	0.509	0.002	0.478	0.066	0.066	0.019	0.002	0.001
Note: Table entries are colored red for values less than 0.33, yellow for values between 0.33 and 0.66 inclusive, and green for values greater than 0.66									

4.2.5 External stations

Table 23 provides the estimated total traffic volume at external stations based on ODOT data and the ATD: again, the large difference in sample size for Vendor B's data can be seen. Table 24 provides the corresponding overall goodness of fit measures. Table 18 suggests mediocre fit to ODOT data for Vendor A's data, but good fit for Vendor B and Vendor C's data.

Table 23: Traffic at external stations – total volume

Data	Traffic volume total
ODOT 2011 (total)	153,859
Non-work	83,026
Work	38,920
Truck	31,913
Vendor A: WTD	69,435
Vendor A: SP	43,399
Vendor B: Personal	528,707
Vendor B: Commercial	3,539,005
Vendor C	77,612

Table 24: Traffic at external stations - overall goodness of fit measures

ODOT data	ATD	Absolute mean error	Root mean square error	R Square
ODOT 2011 (total)	Vendor A: WTD	2113.8	4560.7	0.57
ODOT 2011 (total)	Vendor A: SP	2230.7	4629.6	0.71
ODOT 2011 Non-work + work	Vendor B: Personal	8068.7	26103.0	0.94
ODOT 2011 Truck	Vendor B: Commercial	65823.3	218380.7	0.95
ODOT 2011 Truck	Vendor C	950.8	3073.3	0.99

Figure 51, Figure 52 and Figure 53 compare estimated traffic volumes at external stations based on ODOT data versus ATD at the fifteen external stations with the highest proportional flow totals (based on ODOT totals). The pattern in Figure 51 reflects the mediocre goodness of fit for Vendor A data with ODOT data on all vehicles. Vendor A data processed with both weighted and shortest path-based assignment generally replicates the declining proportional flows from the highest volume external station through the next fourteen stations in order, but not with the monotonically declining pattern indicated by the ODOT data. In contrast, Vendor B data on personal vehicles (Figure 52) and Vendor B and Vendor C data on trucks replicate the monotonically declining pattern evident in the corresponding ODOT data.

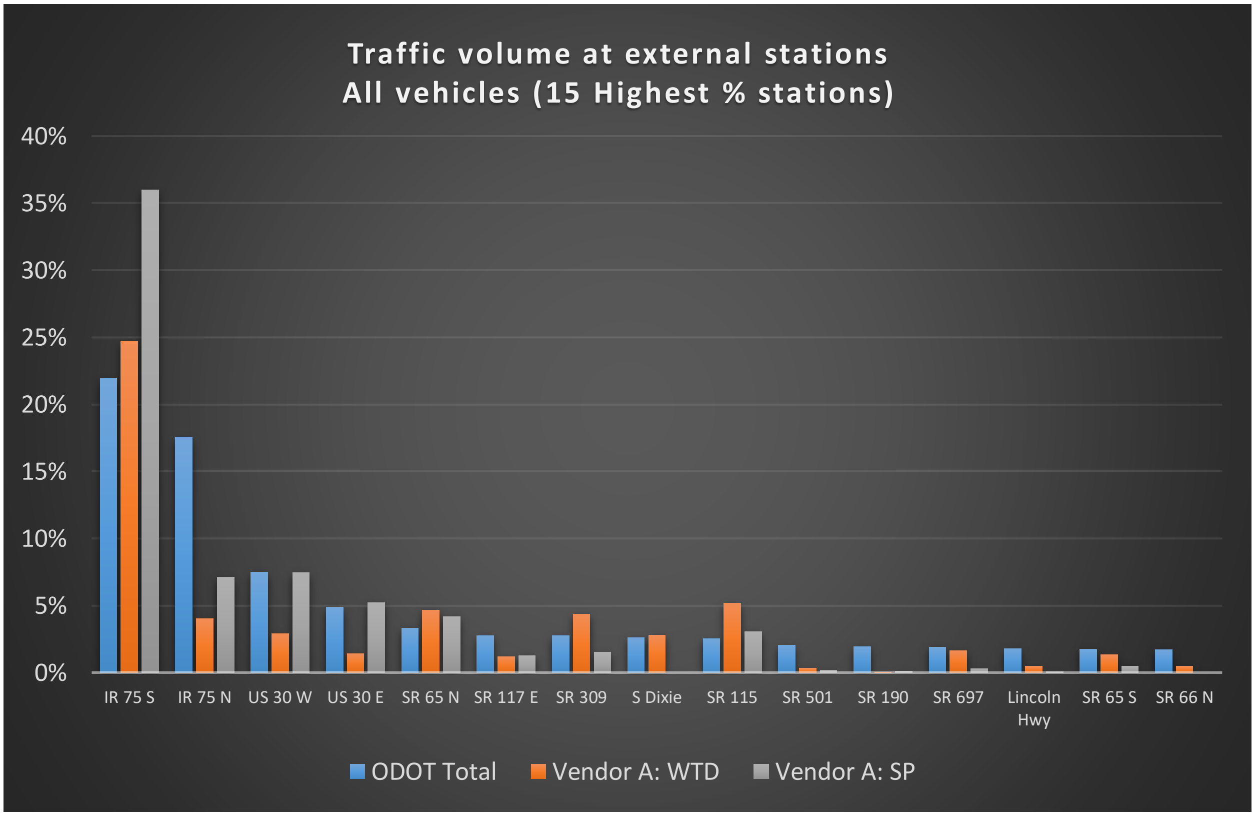


Figure 51: Estimated traffic volumes for all vehicles at high volume external stations - ODOT versus Vendor A

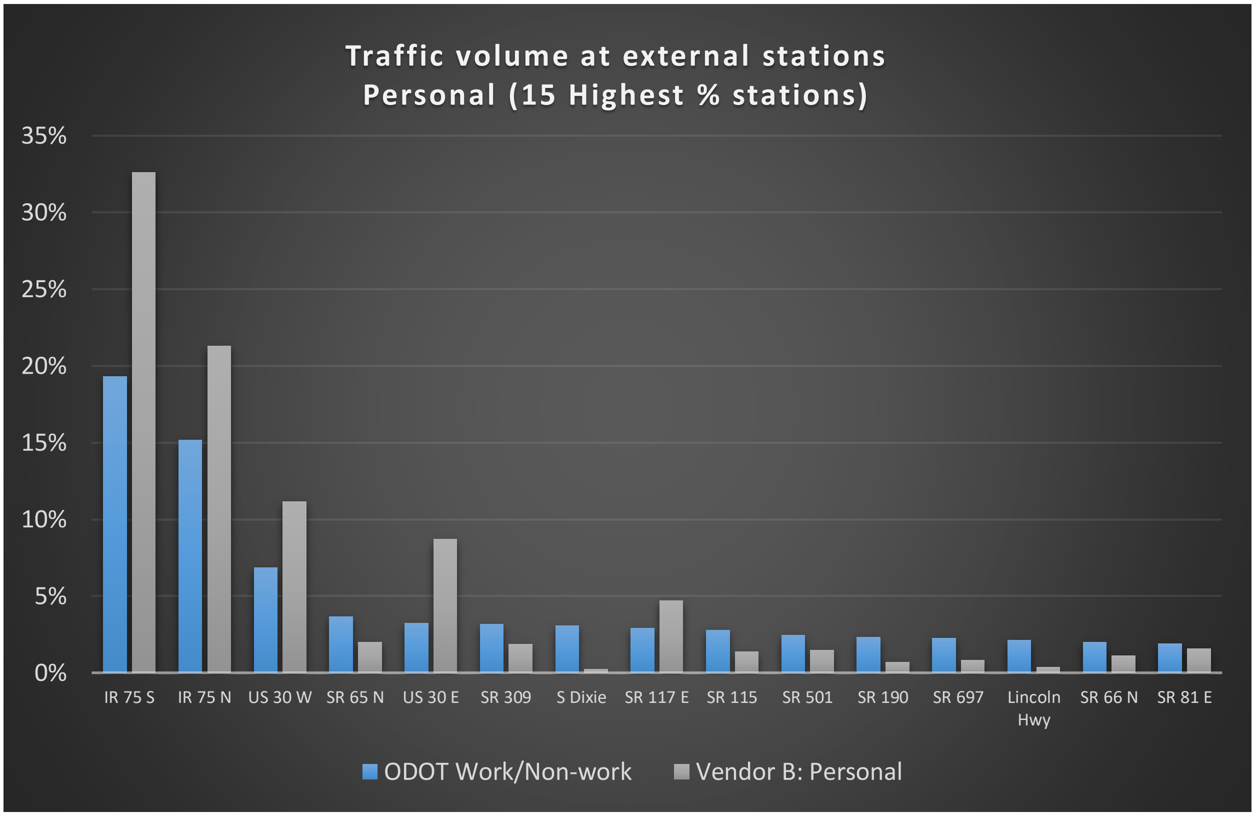


Figure 52: Estimated traffic volume for personal vehicles at high volume external stations - ODOT versus Vendor B

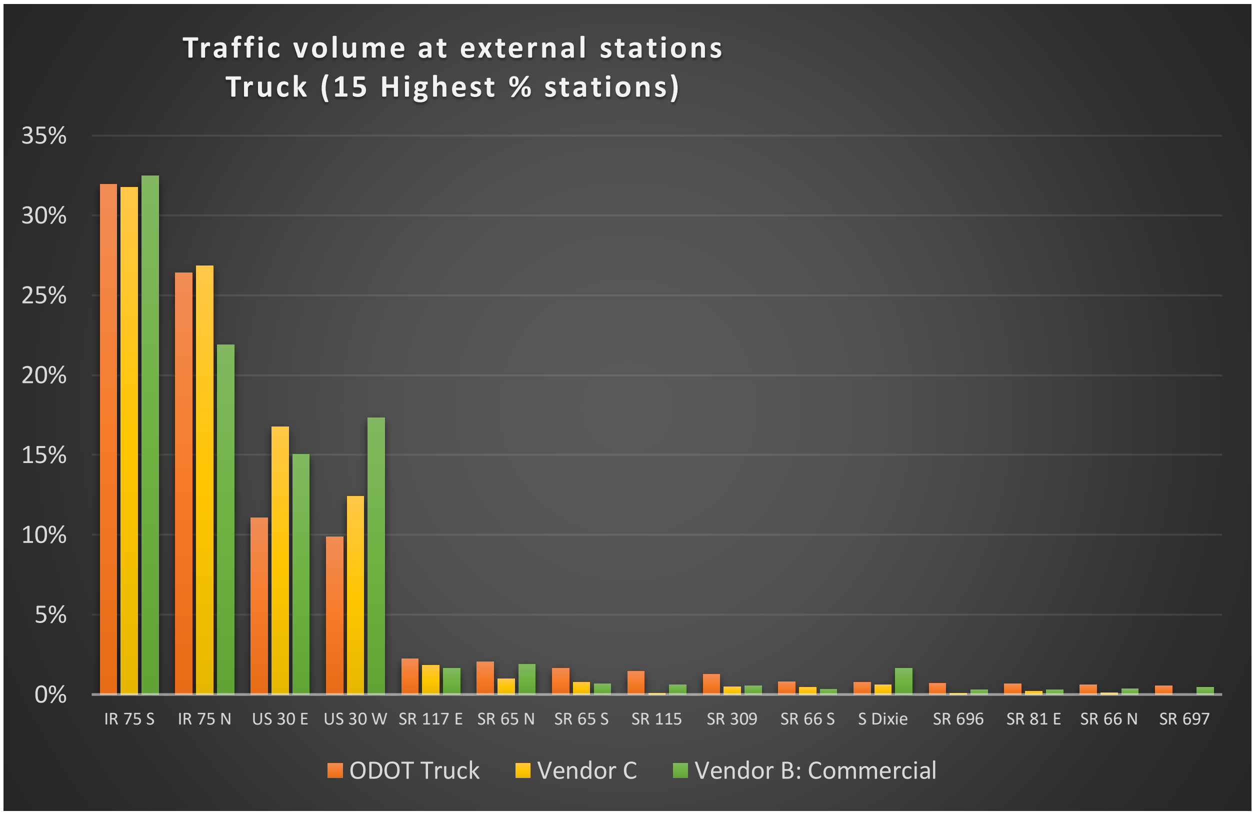


Figure 53: Estimated traffic volume for trucks at high volume external stations - ODOT versus Vendor B and Vendor C

4.2.6 Summary of results

Based on the comparative analysis in the previous sections, the major results are summarized:

1. The ATD has more data with more general spatial coverage of the study area than ODOT data.
2. There is a need to reconcile the different data scales between the ATD and the ODOT data.
3. The ATD fit the ODOT data well for spatially aggregate outcomes such as trip lengths, EE flows and traffic volumes at external stations, although in a relative rather than absolute sense due to the differences in data scales
4. The ATD did not fit the ODOT data well for spatially disaggregate outcomes such as IE/EI flows, TAZ flow totals and trip purpose
5. ATD based on GPS devices fits ODOT data better than ATD based on triangulating cell phone locations.
6. ATD based on cell phones required additional processing to assign flows to external stations. Of the two methods applied, AADT-based weighting and shortest path assignment, shortest path assignment improves fit dramatically.

4.3 Modeling impacts of differences

Archived data products, such as cellular phone and GPS device traces, play an increasingly important role in estimating and validating steps of travel demand models. One important application of these data is in developing external trip models, for trips originating from and/or destined for locations beyond the model area boundaries. In this context, origin-destination matrices constructed by commercial firms replace information collected in and expanded from intercept surveys.

In this analysis, we assign origin-destination matrices from the different providers as an exercise in determining how to work with these data. We identify different procedures for working with data from each provider and comment on differences that lead to different outcomes in the final model.

4.3.1 Methodology

The Ohio Department of Transportation provided the full travel demand model and data input files for the Lima/Allen County Ohio region demand model. We adapted the assignment modules from the Lima model into a free-standing application that assigns vehicle trip matrices by vehicle class (auto and truck) in daily periods.

In this analysis, the internal (II) trip matrices remain fixed and are extracted from a 2011 calibration-year run of the Lima model. The external matrices change with the analysis scenario; these scenarios are as follows:

1. Base
2. Raw Vendor A
3. Scaled Vendor A
4. IPF Vendor A
5. Scaled Vendor B
6. IPF Vendor B

The base scenario applies the external (IE, EE, and EI) trip matrices from the Lima model directly. A description of each other scenario and how its external trip matrices were assembled is given in the following sections.

A “raw” scenario assigns the matrix from the provider directly, only disaggregated among external stations. A “scaled” scenario naively adjusts the raw trip matrix up or down to match the total volume of external flows at the origin based on count data. Finally, an “IPF” scenario applies an iterative proportional fitting algorithm to adjust the raw matrix such that it matches the external counts.

4.3.2 Counts

There is some unavoidable mixing of targets in this analysis; the “Base” scenario demand matrix is a matrix that has been calibrated to 2011 demand and coverage counts. The archived datasets we are comparing against were collected in 2014.

We received from ODOT a set of 53 AADT counts in 2014 at the external stations. Of these, 21 have information on the share of trucks. For those that did not, we imputed a truck share via hot-deck imputation (we selected a random value from a similar station). We did this because the input files to the Lima model include truck volumes at all external stations, and because the IPF process can fail to converge when some marginals are zero.

The 2011 model external productions match reasonably well with the 2014 AADT counts, both in total and at a station-by-station level, as the table below shows. As a note, because neither the count nor the passive archived data discriminate by trip purpose, the “auto” values include both work and non-work productions from the model.

4.3.3 Districts

We created districts for the purpose of aggregate matrix comparison. The Lima model input files did not specify any district aggregation, so we developed the districts shown below. As these are external trips, there are separate districts for each boundary of I-75, US 30, and SR-65, in addition to groups of smaller facilities.

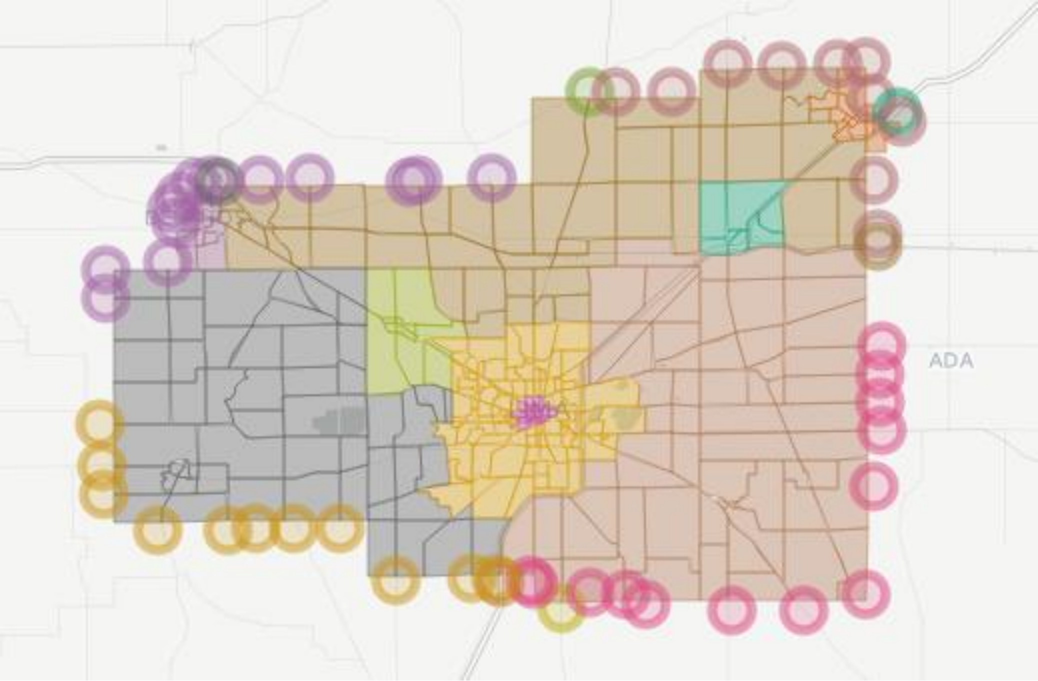


Figure 54 - District aggregations

4.3.4 Vendor A

Vendor A collects phone location data from cellular phone towers. Individuality and market penetration are advantages over other potential methods, but there can be error in spatial or temporal measurements. Vendor A aggregates their data and identifies home and work locations.

Vendor A does not discriminate between passenger and commercial traffic, the raw observed Vendor A flows were split into automobile and truck traffic based on the total truck and auto AADT at the external stations.

The Vendor A matrices are available by time-of-day, but the ODOT AADT counts are (by definition) 24-hour volumes. To scale the provider matrix to match the AADT count, time-of-day factors derived from the provider matrix were applied to the AADT counts. Then all inbound trips were factored to match half the AADT count at the origin. The formula for this is:

$$A'_{ijp} = A_{ijp} * \frac{0.5 * AADT_i * \sum_{ij} A_{ijp} / \sum_{ijp} A_{ijp}}{\sum_{ij} A_{ijp}}$$

Where

A_{ijp} are trips from i to j in period p in vendor matrix A.

- $\sum_x A_{ijp}$is the sum of raw matrix A along dimension x
- A'_{ijp} are the adjusted trips from i to j in period p

Thus at the origin side, the daily inbound vehicles match half the station count, and outbound flows from other stations were relied upon to make up the other half of the AADT. As a note, the total volume of all external-external and internal-external trips must equal half of all two-way station counts to maintain conservation of trips and a daily average travel assumption. Passenger vehicles and trucks were scaled independently.

For the IPF scenario, a period target was derived for each external station by applying the same time-of-day factors as above to half the station count. IPF (or Fratar) was then used to adjust the raw provider matrix to match these counts. This is an improvement to the scaling process, as it attempts to match the counts on the inbound *and* the outbound side; however, it is possible that the method may substantially distort the original matrix, resulting in flows that are more determined by the count values than the original matrix.

As a note, the process failed to converge to a solution within a modest tolerance level. The final iteration results were used with the awareness that there is some error between the input targets and the final output matrix (the maximum error is between 10 and 40 trips at a station in a period).

4.3.5 Vendor B

Vendor B collects in-vehicle GPS data from navigation providers. There is more precise location and time information than from using cellular phone towers, but the penetration rates are lower and potentially biased towards fleet operators. Perhaps because penetration rates are lower, Vendor B scales its data products using seasonal and other adjustment factors; even the “raw” data have been processed to protect confidentiality and the matrices retrieved from them are not proportional to the actual traffic. Because of this, there can be no “raw” Vendor B scenario.

For the scaled and IPF scenarios, the same methods were followed for Vendor B as with Vendor A, with the exception that time-of-day factors were recalculated based on the volumes in Vendor B’s matrices.

4.3.6 Results and Analysis

The table below shows the total daily demand by vehicle type in all scenarios. Unsurprisingly, the scaled scenarios match almost precisely the inbound counts. The Raw A matrix also matches the expected demand closely. The IPF scenarios both have substantially more total demand than the external flows alone should indicate, though curiously neither has nearly as much demand as the Base scenario.

Table 25 - Total daily demand by vehicle type

Scenario	Auto	Truck
Base	107332.59	21314.21
Raw A	66356.45	14455.77
Scaled A	61387.73	17241.77
IPF A	90752.30	26924.73
Scaled B	62130.94	17314.96
IPF B	85792.08	21838.29
11 Externals/2	63220.50	16792.50
15 Counts/2	62166.04	17314.96

Table 26 - Total daily demand by vehicle type and time period

scenario	auto_am	auto_md	auto_pm	auto_nt	auto	truck_am	truck_md	truck_pm	truck_nt	truck
Base	23553.15	19598.82	35213.39	28967.23	107332.6	3401.68	7689.59	5111.47	5111.47	21314.21
Raw A	9848.967	26003.37	14253.79	16250.31	66356.45	2145.601	5664.844	3105.193	3540.136	14455.77
Scaled A	9111.484	24056.26	13186.48	15033.5	61387.73	2559.113	6756.604	3703.643	4222.411	17241.77
IPF A	14134.6	35043.09	19690.52	21884.09	90752.3	4153.048	10456.56	5814.523	6500.6	26924.73
Scaled B	3314.592	11682.21	21254.42	25879.72	62130.94	923.7273	3255.657	5923.288	7212.291	17314.96
IPF B	4778.634	16621.67	29194.93	35196.85	85792.08	1155.812	4180.153	7452.089	9050.241	21838.29

4.3.7 Assignment Validation

It is typical to validate travel model assignment results against AADT or counts at stations distributed throughout the model network. The following tables show the percent RMSE by area type and facility type. In general, the IPF scenarios have RMSE statistics comparable to the existing model, and in fact Vendor B exceeds the model validation on some facility types. The scaled scenarios have poor fit on higher functional class roads and in rural areas; it should also be noted, however, that internal trips dominate lower functional classes and non-rural areas, so the good fit among all scenarios in these areas is not likely a function of the vendor data alone.

Table 27 - Percent RMSE by Facility Type

fclass	Base	IPF A	IPF B	Raw A	Scaled A	Scaled B
Freeway	7.23	17.21	7.1	70.21	55.84	60.18
Expressway	17.59	19.25	11.21	73.37	65.5	58.92
Major Road	37.22	38.62	38.76	40.14	41.7	42.66
Minor Road	50.39	54.46	51.58	56.81	59.16	59.92
Local	90.1	90.04	89.94	90.47	90.28	88.59
Connector	71.5	72.38	72.31	72.56	72.87	72.89

Table 28 - Percent RMSE by Area Type

area_type	Base	IPF A	IPF B	Raw A	Scaled A	Scaled B
Rural	31.8	44.47	32.35	133.47	111.05	117.1
Suburban	46.52	47.94	46.99	47.98	48.94	49.51
Urban	56.53	56.24	57.45	59.04	58.73	59.49
CBD	60.43	63.32	63.45	64.5	65.52	66.36
Outlying BD	25.6	28.48	27.87	29.37	31.74	32.49

NCHRP Report 765 recommends a statistic known as *maximum desirable error*, which uses the ratio of links within a recommended error threshold. This recommendation is based on the observations of numerous studies showing substantial statistical variance in AADT measurements at continuous counting stations and in factoring methodologies. This threshold is shown below, with the error in the modeled flows plotted against the coverage count volumes.

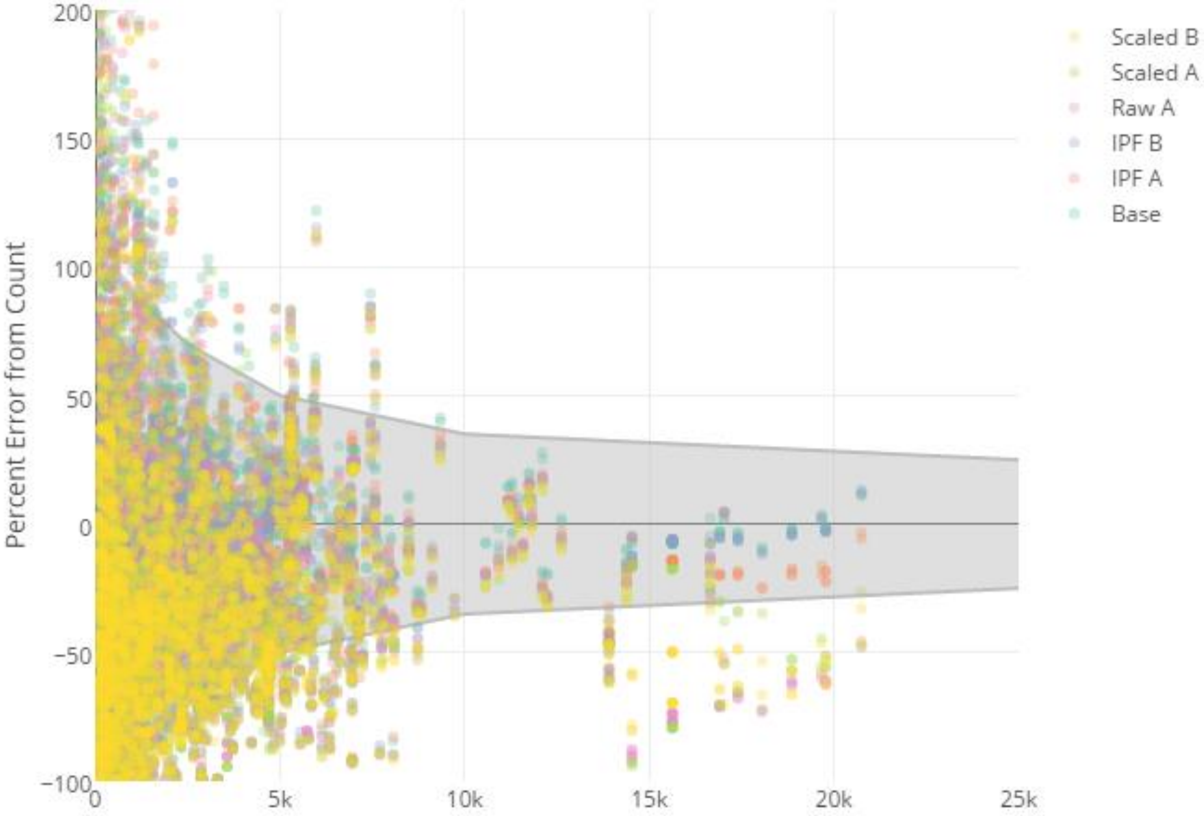


Figure 55 - 2011 Count

4.3.8 Discussions and Recommendations

Archived third-party data products are an increasingly important component of travel models, particularly for external trip models. Working successfully with these products may require some adjustment to historical practices.

First, the IPF process substantially adjusts a provider matrix to match on-the-ground counts at external stations, but it does not completely dictate the shape of the final matrix. This makes having quality count data essential. In this study, truck counts were imputed at external stations where these data were missing, which may have affected the end result. Fewer count external stations with better data may result in a better product.

Along these same lines, archived data products do not effectively distinguish between work and non-work trip purposes, so we did not attempt to compare the assignment results by purpose.

5 FINAL ODOT RECOMMENDATIONS

5.1 Replacing Traditional Cordon Surveys with ATD

The primary objective of this research was to determine if ATD can be used to replace the costly and burdensome traditional cordon survey process in Ohio. Results from this analysis of Lima, Ohio paint a complicated picture where a confident and single recommendation is not possible. The analysis procedures indicated that Vendor B performed the best of the three products. Further, Vendor B has both a personal and commercial product that eases the integration effort. Arguments that prevent a full and confident recommendation of Vendor B (or any ATD product) is that the overall performance was not as strong as hoped and the analysis was limited to a single study area. The original study questions and secondary study questions are as follows:

Does ATD offer any quality or sampling improvements or limitations that enhance or limit traditional travel demand forecasting model performance?

Based on the analysis of Lima, OH, the ATD products do not offer unambiguous quality improvements when compared with ODOT data. Sampling improvements with ATD are evidentially better as they do not share the spatial and temporal limitations of traditional surveys. EE, EI, and IE trip information can be gathered for an extended period of time, a clear advantage over traditional methods. However, there does appear to be bias in the ATD that is difficult to identify given the study parameters. The primary reasons for concern are noted throughout the research discussion but include poor results in comparing the survey station to zone flows, lack of trip purpose details, and product-specific issues. On the positive side, an understanding of the ATD sources and methods allows for the effective use of pre-processing techniques to improve performance. Section 5.2 provide specific details on these methods.

Vendor B performed better than the other solutions potentially due to the fact that their raw data has a better spatial resolution and it is therefore better at defining the exact external station entry and exit points.

What archived cellular data specifications should be applied to maximize value and minimize cost?

The specifications for each product may change based on offerings from each provider, but the following items should be considered when negotiating the purchase of a product:

5.1.1 Vendor A

Vendor A offers a wide variety of specifications. For the specific implementation of Vendor A data as a replacement of traditional cordon surveys and based on the Lima OH analysis, the following specifications are recommended to maximize value:

1. Do not opt for trip type variation – results of trip type comparison were inconclusive, however, comparisons between the survey data and Vendor A's data showed significant differences. Subsequent analysis of large employment center showed low numbers of work related trips by Vendor A. It is possible that Vendor A could improve their capabilities in a location with higher populations, higher market penetration, and improved capability to identify EE, IE and EI trip end locations outside of the study area. The current limited catchment area approach may be limiting their ability to estimate trips and trip types.
2. Do not opt for person type variation – results could not be verified, and the person type provided limited value to the traditional external travel model development.
3. Aggregate TAZs to larger districts – the TAZ level analysis showed significant variation from the survey. Reducing the number of zones reduces the cost and potentially improves the quality of

the data. This also requires that trip disaggregation techniques be applied after the data purchase to assign district level trip information to specific TAZs.

4. Clearly define the catchment areas outside of the study area – large catchment areas along travel corridors will likely provide better data. Due to the nature of Vendor A's methods, it is expected that this is critical to capturing the external travel.

5.1.2 Vendor B

Vendor B offered fewer specification details and appeared to limit options based on their internal structure of their data. The pricing of the product also appeared more fluid, making it more difficult to determine the value of various specifications.

1. Aggregate TAZs to larger districts – the TAZ level analysis showed significant variation from the survey. Reducing the number of zones potentially reduces the cost and potentially improves the quality of the data. This also requires that trip disaggregation techniques be applied after the data purchase to assign district level trip information to specific TAZs.

2. Expand the temporal coverage of the data – This analysis relied on a single month of data, but there are potential, yet unproven, advantages to selecting multiple months or seasonal data collection periods.

3. Provide specific external station location details – Vendor B has good geospatial information that can be used to match to specific external stations. However, lightly traveled stations may show more bias as the expected raw data source penetration is limited.

4. Identify specific truck weigh stations and truck stops – Analysis of Vendor B commercial data in Lima suggested that their trip end identification techniques mis-identified weigh stations and truck stops as an origin or destination. While these vehicles did actually stop at those locations and this is significant for modeling, a full understanding of where a truck trip was going was limited.

5.1.3 Vendor C

Vendor C offered fewer specification details and significant privacy protection restrictions on their data. Their data do not come fully processed into origin-destination details but are simply traces of travel details using a time-limited window. Spatially, the travel trace is provided as a sequence of zone-IDs along with vehicle speed and timestamp. Vendor C was also the most inexpensive source due to its raw form and the nature of the organization providing the data. The following specifications are suggested if ordering Vendor C data:

1. Provide TAZs – Vendor C will use zones specific to your study area. As a default they use US Census block groups.

2. Provide external catchment areas as TAZs – Since Vendor C uses US Census block groups as default zone ID for identifying the location of trip details, there is a challenge in determining actual travel roads and entry/exit points of the study area. Providing a catchment area extending out from each external station eliminated this problem.

How can archived travel data be applied in traditional travel demand forecasting models to make maximum use of its strengths and minimize the impact of its limitations?

It is important to understand that all data sources have limitations and contain margins of error. One can consider ATD and surveys as two ends of a spectrum, where travel surveys can provide maximum detail on individual trips and ATD can provide a much deeper sample of trips from a global perspective.

An external trip model that fully utilizes the strengths of ATD may look different from current models based on intercept survey data. Because ATD providers' purpose imputation can be weak and underdeveloped, (a finding of this and other studies), modeling internal/external trip attractions by purpose may not be possible or desired. Similarly, given the finding that larger districts and catchment areas result in more coherent results, it is prudent to eliminate external stations with very small flows (less than approximately 1k AADT), particularly if there are many such roads on the same side of the model.

If it is essential that an internal/external trip model capture trip purpose, then ATD alone may not be sufficient. On the other hand, current trip attraction models estimated on survey data typically have large standard errors and low predictive power because the variables available to predict trips (floor space, jobs) are only weakly correlated with trip making.

5.2 Techniques to maximize archived data value

5.2.1 Normalize values to traffic counts

All of the ATD products provided raw trip counts that varied widely in meaning. Vendor A is simply based on observed trips for the period of time specified in the product, but it is not known if any other adjustments are made. Vendor B uses a more complicated method of trip count estimation as documented in their delivered data product. Vendor C uses simple counts from within their database for the given period of travel (no adjustments) but their raw sources are limited. Use of each product should be normalized to high quality vehicle classification counts at all external stations desired for the model. This can be done in a number of ways, including naïve scaling and iterative proportional fitting. We have written an R package available on GitHub that contains functions for both methods.

5.2.2 Zonal aggregation

As noted in the analysis, the evaluation of EI trip flows between external stations and TAZs indicated a poor correlation with the trip flows identified through ODOT data. Explanations for this mismatch are varied and not entirely understood by the research community. Some reasons could include; bias in the demographic profiles of raw data sources, major trip attractors/producers close to the boundaries of TAZs, poor spatial resolution of archived data trip ends, imperfect trip end determination algorithms, and changes in travel patterns in the community. Given that ATD products are under constant improvement/change due to technology advances or changes in source data, the specific set of biases impacting trip end intensity and location errors are likely to remain unknown or difficult to identify.

One method for minimizing the impact of these errors is to aggregate the TAZs into larger zones. This has the potential for balancing TAZ trip ends counts between zones with higher than expected values and zones with lower than expected values. Modeling techniques could then be used to disaggregate (or impute) trips back to specific TAZs based on known demographic data such as population, employment and landuse.

This approach was tested on Vendor A and Vendor B personal travel data. TAZs within the study area were aggregated to "districts" which were already coded into the data provided by ODOT. One exception was made in that the Lima central business district was expanded. The EI trip ends were then aggregated by district for the 2008 survey data, the most recent model base year estimate, and all of the ATD products. **Table 29** shows the RMSE and R^2 values between Vendor A / Vendor B and the 2008 survey. These values show improvement from the original EI analysis with the raw data. Similarly, **Table 30** shows improvement when the same values are compared with the aggregated model data. These comparisons, while improved, still suggest that there is significant bias with each ATD.

Table 29 - Comparison of EI trip end density to 2008 survey (personal travel)

	Vendor A	Vendor B
RMSE	0.142	0.137
R-SQ	0.324	0.462

Table 30 - Comparison of EI trip end density to current model (personal travel)

	Vendor A	Vendor B
RMSE	0.121	0.110
R-SQ	0.666	0.798

Table 31 shows the spatial patterns evident in the aggregated data. The figures indicate a spatial consistency between the survey, model, Vendor A and Vendor B datasets. It should be noted that the ATD products compare better with the model baseline EI trip densities than they do with the survey data.

Table 31 - Maps of EI trips end density by district (personal travel)

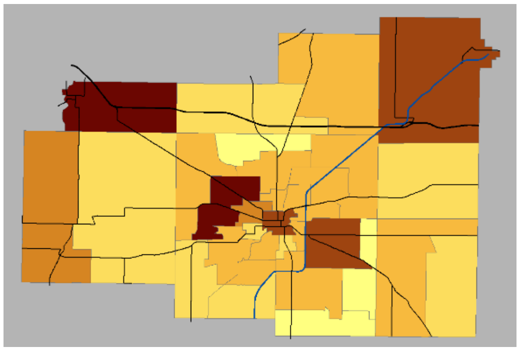


Figure 56 - 2008 Survey EI trips by district

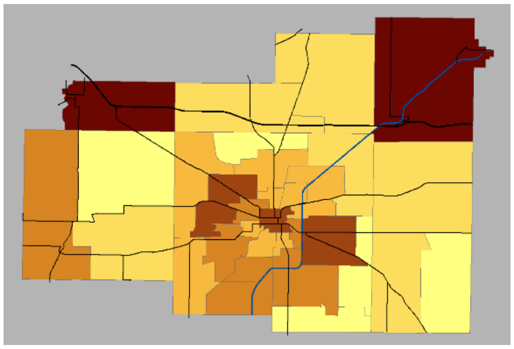
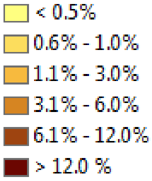


Figure 57 - Model EI trips by district



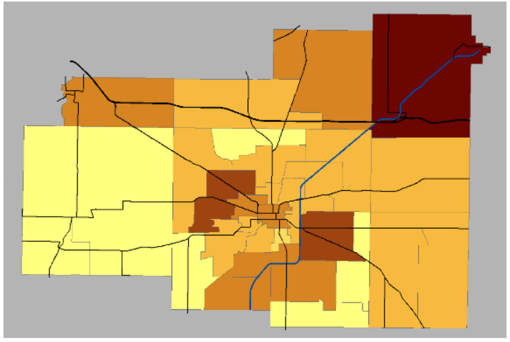


Figure 58 – Vendor A EI trips by district

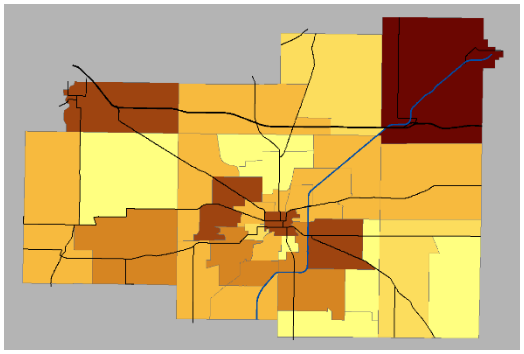


Figure 59 – Vendor B (personal) EI trips by district

The same district aggregation approach was applied to the commercial travel datasets. The RMSE and R^2 values did not indicate substantial improvement over the original TAZ-level comparison (see **Table**

32 and **Table 33**). Note that in this case, both ATD products compared better to the survey data instead of the model.

Table 32 - Comparison of EI trip end density to 2008 survey (commercial travel)

	Vendor B	Vendor C
RMSE	0.148	0.192
R-SQ	0.477	0.250

Table 33 - Comparison of EI trip end density to current model (commercial travel)

	Vendor B	Vendor C
RMSE	0.164	0.198
R-SQ	0.192	0.193

Table 34 shows the spatial patterns in the commercial travel for the survey, base year model, Vendor B, and Vendor C. Significant variation exists on the interior districts amongst the products, particularly surrounding the Lima area.

Table 34 - Maps of EI trips end density by district (commercial travel)

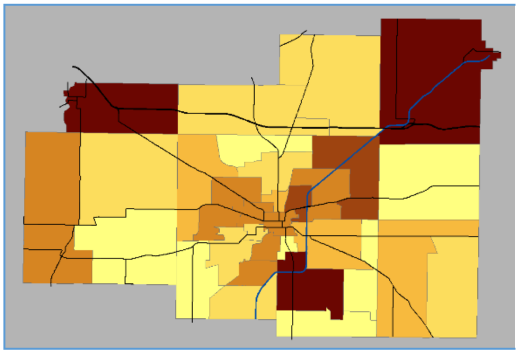


Figure 60 - 2008 survey truck EI trip end density

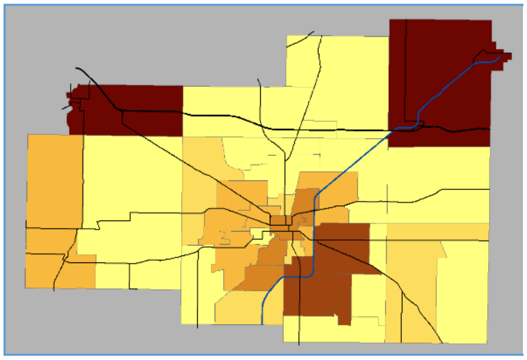


Figure 61 - Model truck EI trip end density

🟨 < 0.5%
🟨 0.6% - 1.0%
🟧 1.1% - 3.0%
🟧 3.1% - 6.0%
🟫 6.1% - 12.0%
🟥 > 12.0 %

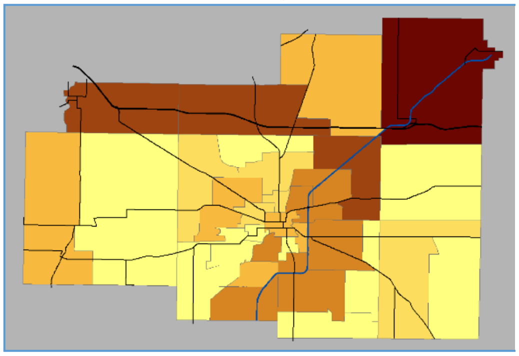


Figure 62 – Vendor B commercial EI trips end density

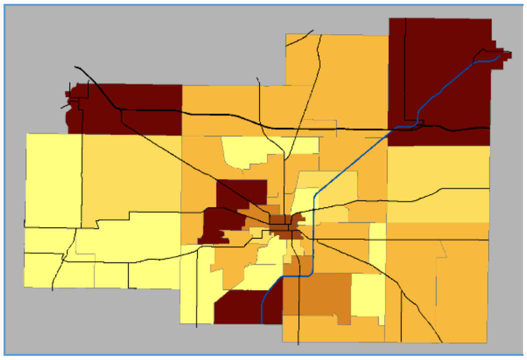


Figure 63 –Vendor C EI trip end density

5.2.3 External catchment methods

The external catchment zones defined by Vendor A in this study were fairly large and covered multiple entry/exit points into the study area. Isolating the specific EE, EI, and IE entry/exit points therefore required a method to assign trip percentages to each entry/exit point. Initially, and based on past research recommendations, these trips were allocated based on the ADT of each station. This resulted in errors that included the misassignment of EE trips between adjacent catchment zones and illogical trip paths for short trips. The research team therefore devised a shortest path method for assigning entry/exit points that proved to eliminate these errors. The solution treated each catchment area as a TAZ with a centroid and centroid connectors that loosely resembled the road network and connected each entry/exit point. Connections between adjacent catchment zone centroid were also created. Trips between zones were then forced to one of these entry/exit points based on a shortest path between their origin and destination. This eliminated the adjacent trips as these trips never entered the study area. Longer trips were assigned to those roads with better connectivity. Shorter trips were assigned to entry/exit points more logical for their trip path. This approach significantly improved Vendor A's comparison.

5.2.4 Truck Data

As previously discussed, the identification of truck trip origins and destinations is challenging when attempting to define these through automated data procedures. Typically, the automatic trip end identification methods define a threshold stop time and possibly combined with a travel distance. Trucks sometimes have to stop at weigh stations (see **Figure 64**) or truck stops (**Figure 65**) and sometimes these stops can be longer than a few minutes. Further, within a destination, a truck can stop and then move around the large industrial site multiple times potentially resulting in false trip ends.

If using Vendor C, these techniques need to be assessed and implemented on the raw data to extract estimates of true origins and destination. For Vendor B commercial data, ODOT should simply communicate this issue to the vendor and then evaluate the results before applying them to the model.

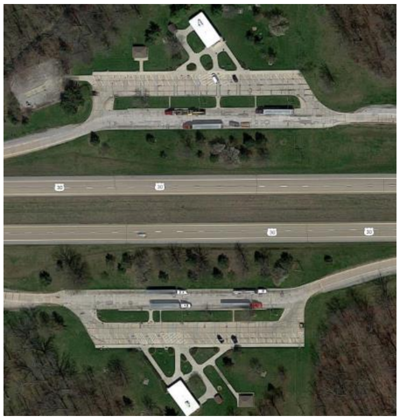


Figure 64 - US 30 Weigh Station

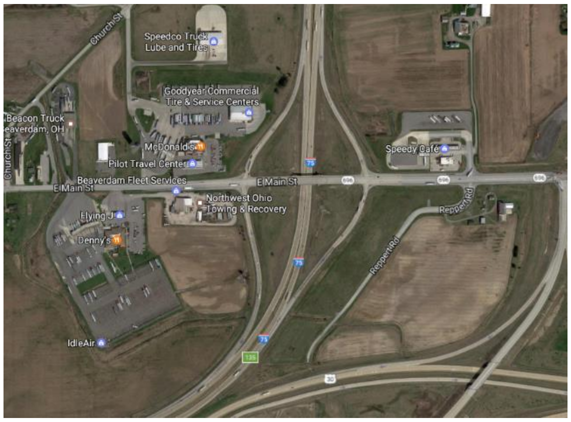


Figure 65 - US 30 and I-75 Truck Stop Area

6 REFERENCES

Alexander, L., Jiang, S., Murga, M., & González, M. C. (2015). Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies*, *58*, 240-250.

Bierlaire, M. (2002). The total demand scale: a new measure of quality for static and dynamic origin–destination trip tables. *Transportation Research Part B: Methodological*, *36*(9), 837-850.

Blogg, M., Semler, C., Hingorani, M., & Troutbeck, R. (2010, September). Travel time and origin-destination data collection using Bluetooth MAC address readers. In *Australasian Transport Research Forum* (pp. 1-15).

Caceres, N., Wideberg, J. P., & Benitez, F. G. (2007). Deriving origin destination data from a mobile phone network. *Intelligent Transport Systems, IET*, *1*(1), 15-26.
Caceres, N., Romero, L. M., Benitez, F. G., & del Castillo, J. M. (2012). Traffic flow estimation models using cellular phone data. *Intelligent Transportation Systems, IEEE Transactions on*, *13*(3), 1430-1441.

Calabrese, F., Di Lorenzo, G., Liu, L., & Ratti, C. (2011). Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Computing*, *10*(4), 0036-44.

Calabrese, F., Di Lorenzo, G., Ferreira, J., & Ratti, C. (2013). Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation research part C: emerging technologies*, *26*, 301-313.

Chen, A., Chootinan, P., Ryu, S., & Wong, S. C. (2012). Quality Measures of Origin-Destination Trip Table Estimated from Traffic Counts: Review and New Generalized Demand Scale Measure. *Journal of Transportation Engineering*, *138*(11), 1340-1349.

Chitturi, M., Shaw, J., Campbell, J., & Noyce, D. (2014). Validation of origin-destination data from Bluetooth reidentification and aerial observation. *Transportation Research Record: Journal of the Transportation Research Board*, (2430), 116-123.

Chow, C., & Mokbel, M. F. (2011). Privacy of spatial trajectories. In Y. Zheng & Y. Zhou (Eds.), *Computing with spatial trajectories* (pp. 109–142). Springer.

Djukic, T., Barceló, J., Bullejos, M., Montero Mercadé, L., Cipriani, E., van Lint, H., & Hoogendoorn, S. (2015). Advanced traffic data for dynamic OD demand estimation: The state of the art and benchmark study. In *TRB 94th Annual Meeting Compendium of Papers* (pp. 1-16).

Gan, L., Yang, H., & Wong, S. C. (2005). Traffic counting location and error bound in origin-destination matrix estimation problems. *Journal of Transportation Engineering*, *131*(7), 524-534.

Hard, E., Chigoy, B., Songchitruksa, P., Farnsworth. S., & Borchardt, D. (2014) Comparison of Cell, GPS, and Bluetooth derived external data. *Travel Survey Methods (ABJ 40) Committee Meeting.*

Iqbal, M. S., Choudhury, C. F., Wang, P., & González, M. C. (2014). Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, *40*, 63-74.

Kwan, M. P. (2016). Algorithmic Geographies: Big Data, Algorithmic Uncertainty, and the Production of Geographic Knowledge. *Annals of the American Association of Geographers*, *106*(2), 274-282.

Larijani, A. N., Olteanu-Raimond, A. M., Perret, J., Brédif, M., & Ziemlicki, C. (2015). Investigating the mobile phone data to estimate the origin destination flow and analysis; case study: Paris region. *Transportation Research Procedia*, *6*, 64-78.

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, *62*(318), 399-402.

Liu, F., Janssens, D., Cui, J., Wang, Y., Wets, G., Cools, M. (2014). Building a validation measure for activity-based transportation models based on mobile phone data. *Expert Systems with Applications* 41(14), 6174-6189.

Lu, Y., & Liu, Y. (2012). Pervasive location acquisition technologies: Opportunities and challenges for geospatial studies. *Computers, Environment and Urban Systems*, *36*(2), 105-108.

Ratti, C., Frenchman, D., Pulselli, R. M., & Williams, S. (2006) Mobile landscapes: using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design* 33.5 (2006): 727-748.

Smoreda, Z., Olteanu-Raimond A. M., & Couronné T. (2013). Spatiotemporal data from mobile phones for personal mobility assessment, In Zmud J, Lee-Gosselin M, Carrasco JA, Munizaga MA (eds), *Transport Survey Methods: Best Practice for Decision Making*, Emerald Group Publishing, London.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 557-570.

Transport for London. (2010). *Traffic Modelling Guidelines Version 3.0*, London, UK.

Van Vliet, D. (2013). *Saturn Software User's Manual (v11.2)*, Epsom, Surrey, UK.

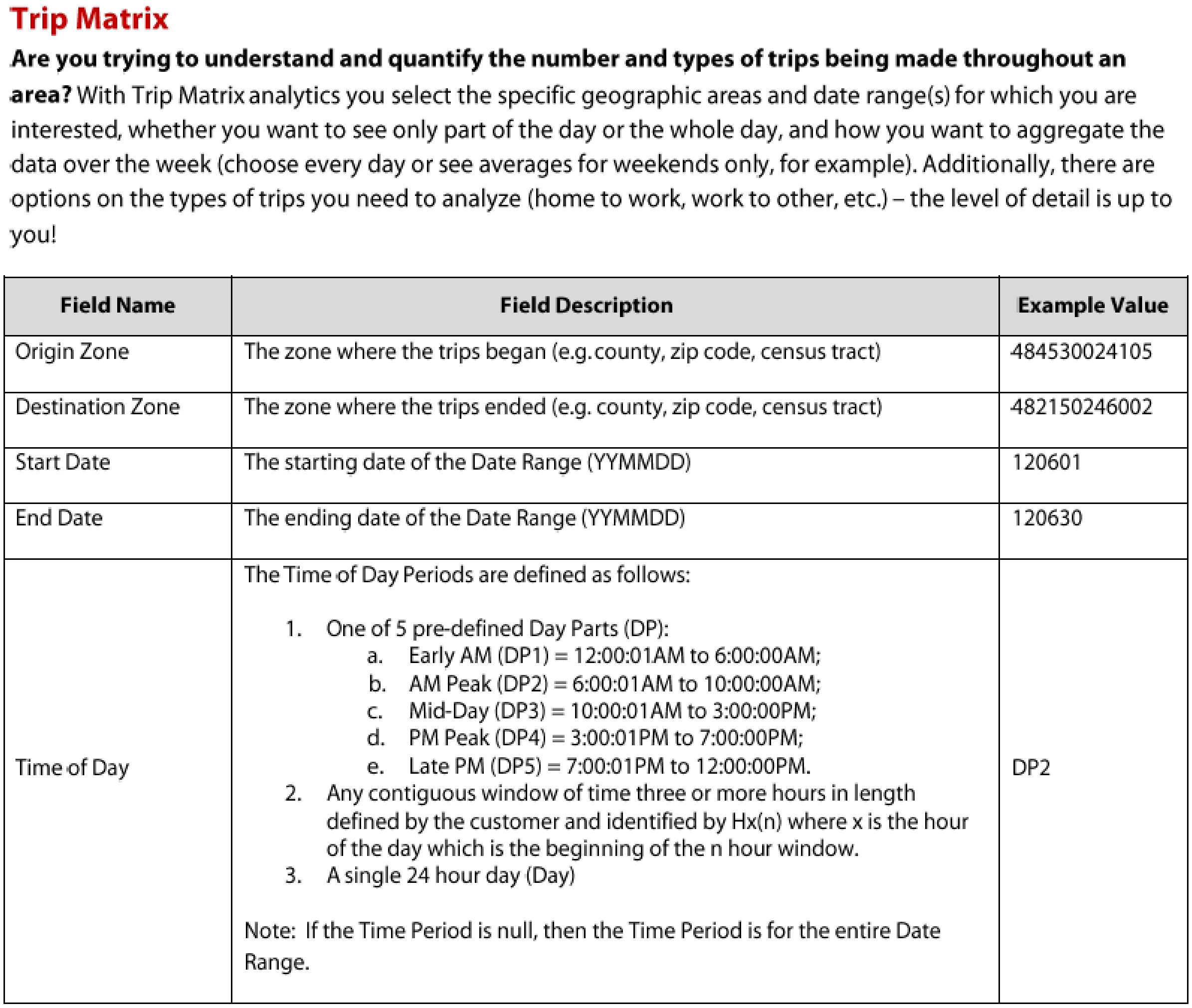
Yang, H., Iida, Y., & Sasaki, T. (1991). An analysis of the reliability of an origin-destination trip matrix estimated from traffic counts. *Transportation Research Part B: Methodological*, *25*(5), 351-363.

Yin, L., Wang, Q., Shaw, S. L., Fang, Z., Hu, J., Tao, Y., & Wang, W. (2015). Re-identification risk versus data utility for aggregated mobility research using mobile phone location data. *PloS one*, 10(10), e0140589.

Zhang, Y., Qin, X., Dong, S., & Ran, B. (2010). Daily OD matrix estimation using cellular probe data. In *89th Annual Meeting Transportation Research Board*.

7 APPENDIX

7.1 Vendor A: ReadMe_TripMatrixAttributes.pdf

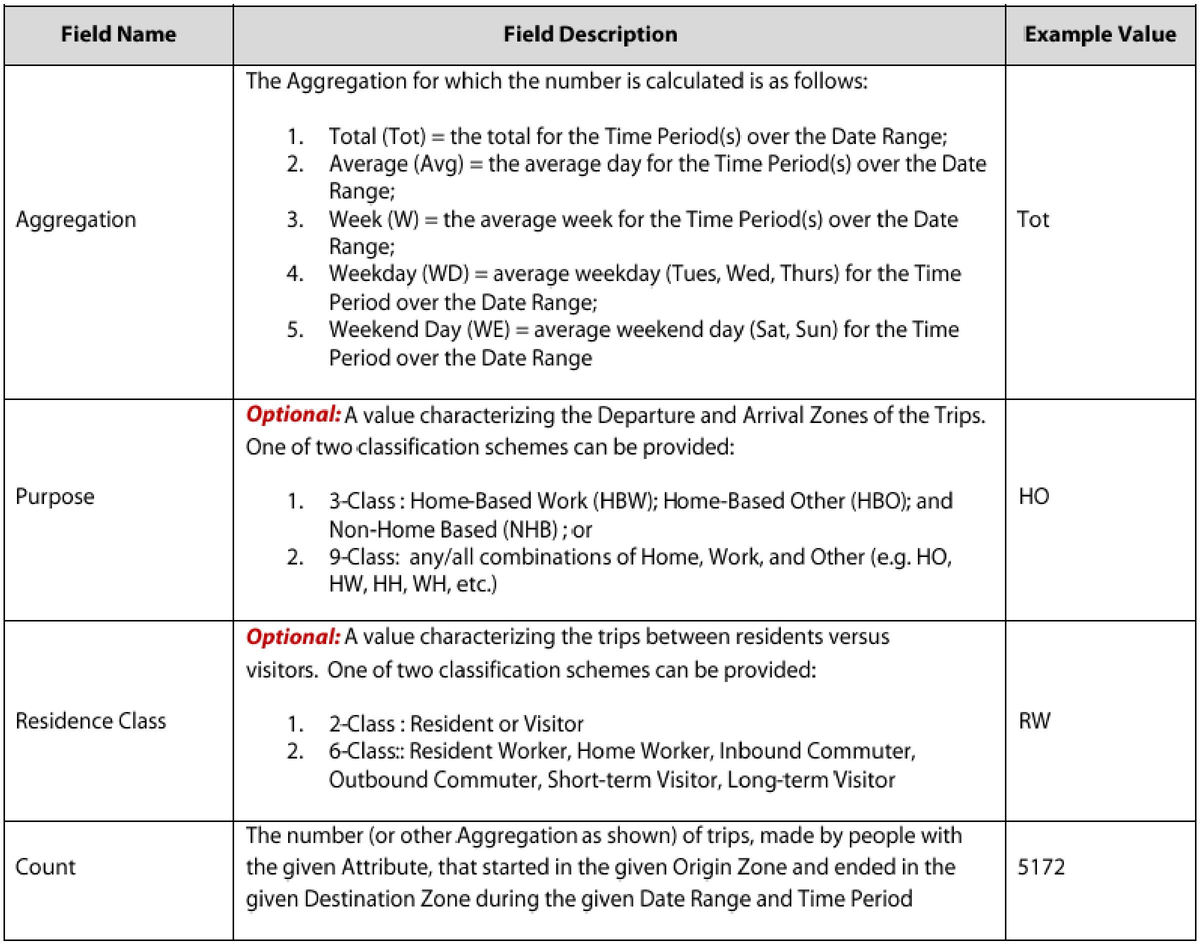


Trip Matrix

Are you trying to understand and quantify the number and types of trips being made throughout an area? With Trip Matrix analytics you select the specific geographic areas and date range(s) for which you are interested, whether you want to see only part of the day or the whole day, and how you want to aggregate the data over the week (choose every day or see averages for weekends only, for example). Additionally, there are options on the types of trips you need to analyze (home to work, work to other, etc.) – the level of detail is up to you!

Field Name	Field Description	Example Value
Origin Zone	The zone where the trips began (e.g.county, zip code, census tract)	484530024105
Destination Zone	The zone where the trips ended (e.g. county, zip code, census tract)	482150246002
Start Date	The starting date of the Date Range (YYMMDD)	120601
End Date	The ending date of the Date Range (YYMMDD)	120630

Time of Day	The Time of Day Periods are defined as follows:
	1. One of 5 pre-defined Day Parts (DP):
	a. Early AM (DP1) = 12:00:01AM to 6:00:00AM;
	b. AM Peak (DP2) = 6:00:01AM to 10:00:00AM;
	c. Mid-Day (DP3) = 10:00:01AM to 3:00:00PM;
	d. PM Peak (DP4) = 3:00:01PM to 7:00:00PM;
	e. Late PM (DP5) = 7:00:01PM to 12:00:00PM.
	2. Any contiguous window of time three or more hours in length defined by the customer and identified by Hx(n) where x is the hour of the day which is the beginning of the n hour window.
	3. A single 24 hour day (Day)
	Note: If the Time Period is null, then the Time Period is for the entire Date Range.
		DP2



Field Name	Field Description	Example Value
Aggregation	The Aggregation for which the number is calculated is as follows: 1. Total (Tot) = the total for the Time Period(s) over the Date Range; 2. Average (Avg) = the average day for the Time Period(s) over the Date Range; 3. Week (W) = the average week for the Time Period(s) over the Date Range; 4. Weekday (WD) = average weekday (Tues, Wed, Thurs) for the Time Period over the Date Range; 5. Weekend Day (WE) = average weekend day (Sat, Sun) for the Time Period over the Date Range	Tot
Purpose	***Optional:*** A value characterizing the Departure and Arrival Zones of the Trips. One of two classification schemes can be provided: 1. 3-Class : Home-Based Work (HBW); Home-Based Other (HBO); and Non-Home Based (NHB) ; or 2. 9-Class: any/all combinations of Home, Work, and Other (e.g. HO, HW, HH, WH, etc.)	HO
Residence Class	***Optional:*** A value characterizing the trips between residents versus visitors. One of two classification schemes can be provided: 1. 2-Class : Resident or Visitor 2. 6-Class:: Resident Worker, Home Worker, Inbound Commuter, Outbound Commuter, Short-term Visitor, Long-term Visitor	RW
Count	The number (or other Aggregation as shown) of trips, made by people with the given Attribute, that started in the given Origin Zone and ended in the given Destination Zone during the given Date Range and Time Period	5172

7.2 Vendor A: ReadMe_AgeKey.docx

The trips made by different age groups are presented in the following fields in the attached dataset.

Age Key	Description
male_Under_5_years	Male less than 5 years
male_5_to_9_years	Male between 5 and 9 years
male_10_to_14_years	Male between 10 and 14 years
male_15_to_17_years	Male between 15 and 17 years
male_18_and_19_years	Male between 18 and 19 years
male_20_years	Male 20 years old
male_21_years	Male 21 years old
male_22_to_24_years	Male between 22 and 24 years
male_25_to_29_years	Male between 25 and 29 years
male_30_to_34_years	Male between 30 and 34 years
male_35_to_39_years	Male between 35 and 39 years
male_40_to_44_years	Male between 40 and 44 years
male_45_to_49_years	Male between 45 and 49 years
male_50_to_54_years	Male between 50 and 54 years
male_55_to_59_years	Male between 55 and 59 years
male_60_and_61_years	Male between 60 and 61 years
male_62_to_64_years	Male between 62 and 64 years
male_65_and_66_years	Male between 65 and 66 years
male_67_to_69_years	Male between 67 and 69 years
male_70_to_74_years	Male between 70 and 74 years
male_75_to_79_years	Male between 75 and 79 years
male_80_to_84_years	Male between 80 and 84 years
male_85_years_and_over	Male greater than 85 years
female_Under_5_years	Female less than 5 years
female_5_to_9_years	Female between 5 and 9 years
female_10_to_14_years	Female between 10 and 14 years
female_15_to_17_years	Female between 15 and 17 years
female_18_and_19_years	Female between 18 and 19 years
female_20_years	Female 20 years old
female_21_years	Female 21 years old
female22_to_24_years	Female between 22 and 24 years
female25_to_29_years	Female between 25 and 29 years
female_30_to_34_years	Female between 30 and 34 years
female_35_to_39_years	Female between 35 and 39 years
female_40_to_44_years	Female between 40 and 44 years
female_45_to_49_years	Female between 45 and 49 years
female_50_to_54_years	Female between 50 and 54 years
female_55_to_59_years	Female between 55 and 59 years
female_60_and_61_years	Female between 60 and 61 years
female_62_to_64_years	Female between 62 and 64 years
female_65_and_66_years	Female between 65 and 66 years
female_67_to_69_years	Female between 67 and 69 years
female_70_to_74_years	Female between 70 and 74 years
female_75_to_79_years	Female between 75 and 79 years
female_80_to_84_years	Female between 80 and 84 years
female_85_years_and_over	Female greater than 85 years

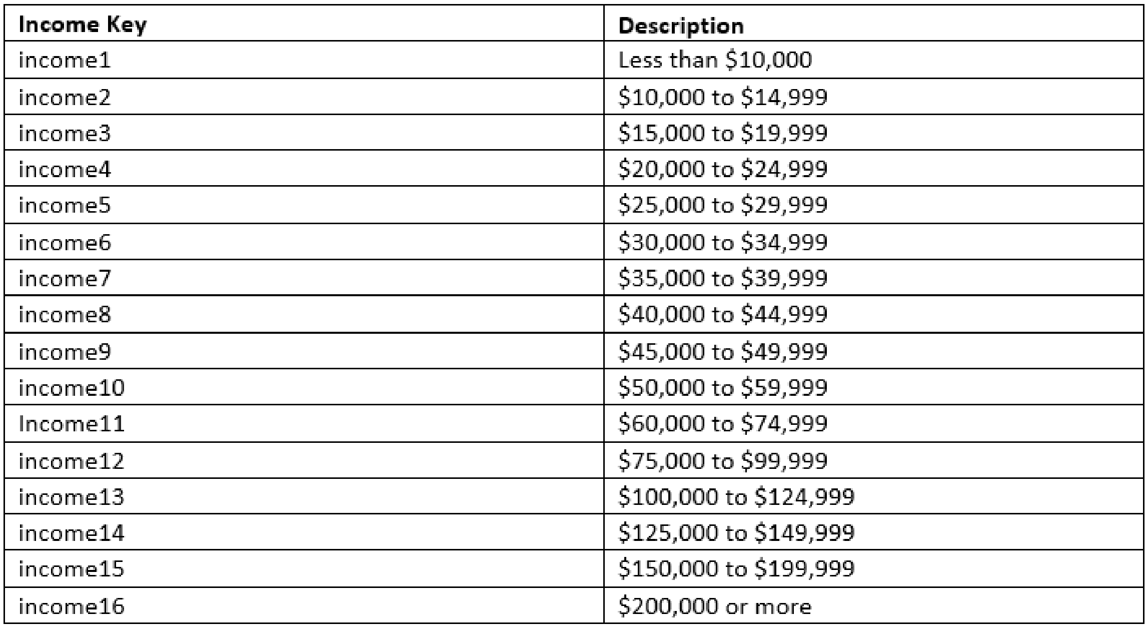
7.3 Vendor A: ReadMe_AutoKey.docx

The trips made by different auto ownership groups are presented in the following fields in the attached dataset.

Auto Key	Description
own_0_vehicle_available	0 vehicles available owned households
own_1_vehicle_available	1 vehicles available owned households
own_2_vehicles_available	2 vehicles available owned households
own_3_vehicles_available	3 vehicles available owned households
own_4_vehicles_available	4 vehicles available owned households
own_5_or_more_vehicles_available	over 5 vehicles available owned households
rent_0_vehicle_available	0 vehicles available rented households
rent_1_vehicle_available	1 vehicles available rented households
rent_2_vehicles_available	2 vehicles available rented households
rent_3_vehicles_available	3 vehicles available rented households
rent_4_vehicles_available	4 vehicles available rented households
rent_5_or_more_vehicles_available	over 5 vehicles available rented households

7.4 Vendor A: ReadMe_IncomeKey.docx

The trips made by different income groups are presented in the following fields in the attached dataset.



Income Key	Description
income1	Less than $10,000
income2	$10,000 to $14,999
income3	$15,000 to $19,999
income4	$20,000 to $24,999
income5	$25,000 to $29,999
income6	$30,000 to $34,999
income7	$35,000 to $39,999
income8	$40,000 to $44,999
income9	$45,000 to $49,999
income10	$50,000 to $59,999
Income11	$60,000 to $74,999
income12	$75,000 to $99,999
income13	$100,000 to $124,999
income14	$125,000 to $149,999
income15	$150,000 to $199,999
income16	$200,000 or more

7.5 Vendor B: Project_OD.txt

Project: Lima-OD

Created by: joan.lim+admin@Vendor Bdata.com

```
Created on: 2016-03-02
Organization: Westat

Project Type: O-D Analysis

Data Period: 2015:[5, 6]
Trip Type: Locked to Route

Day Type:
0: Average Day (M-Su)
1: Average Weekday (M-Th)
2: Average Weekend Day (Sa-Su)

Day Part:
0: All Day (12am-12am)
1: Early AM (12am-6am)
2: Peak AM (6am-10am)
3: Mid-Day (10am-3pm)
4: Peak PM (3pm-7pm)
5: Late PM (7pm-12am)

Commercial Vehicle Results by Weight Class: Disabled
Device Ping Rate for Personal Trips: All
Device Ping Rate for Commercial Trips: All


Metrics Version: R17-M20
Data Months using Trips v13: None
Data Months using Trips v15: 2015:[5, 6]

Copyright © 2011 - 2016, Vendor B Data, Inc. All rights reserved.
```

7.6 Vendor B: README_OD.txt

```
This folder contains Metrics about the Origin-Destination trips between the Zones of the named Project.


Terms
=====
Origin Zone: For this Project, trips were analyzed that started in or initially passed through any of the Origin Zones.
Destination Zone: For the Project, trips were analyzed that ended in or passed through any of the Destination Zones after starting in or passing through an Origin Zone.


Files
=====
Project_OD.txt
==============
This file lists information about the Project as a whole, including the full Project name, organization and user name that created the Project, and the Data Period for the Project.
```

zones.csv
=========
This file contains information about the Zones used in this Project.

- Zone Type: Indicates if the Zone is an Origin or Destination Zone for this Project.
- Zone ID: Numeric ID for the Zone. This is the 'id' from the input shapefile, if provided by the user.
- Zone Name: Name for the Zone. This is the 'name' from in the input shapefile, if provided by the user.
- Zone Direction(degrees): This refers to the direction in which trips pass-through the Zone. Values are provided in degrees from 0 to 359, where 0 is due north, 90 is east, 180 is due south, etc. When creating Metrics, a range of -20/+20 degrees is applied to the Zone Direction value. A value of "Null" refers to no direction filter and therefore all trips that pass-through the Zone will be used. Note: this attribute is only relevant for Zones where "Is Pass-Through" is set to "Yes".
- Zone is Pass-Through: A "Yes" value indicates trips are expected to pass through the Zone and that designation was used when creating the Metrics. This indicates the Zone is likely a road segment, but not necessarily. A "No" value indicates trips are expected to start or end in the Zone and that designation was used when creating the Metrics.

od_personal.csv & od_commercial.csv
===================================
These files contain the OD Metrics for Personal or Commercial trips.

- Vehicle Type: Type of vehicle analyzed with values of 'Personal' or 'Commercial'.
- Origin Zone ID: Numeric ID for the Origin Zone. This is the 'id' from the input shapefile, if provided by the user.
- Origin Zone Name: Name for the Origin Zone. This is the 'name' from in the input shapefile, if provided by the user.
- Origin Zone Is Pass-Through: "Yes" value indicates that only trips passing through the Origin Zone are represented in the frequency. "No" value indicates that only trips that start in the Origin Zone are represented in the frequency values.
- Origin Zone Direction: This refers to the direction in which trips pass-through the Origin Zone. Values are provided in degrees from 0 to 359, where 0 is due north, 90 is east, 180 is due south, etc. When creating Metrics, a range of -20/+20 degrees is applied to the Origin Zone Direction value. A value of "Null" refers to no direction filter and therefore all trips that pass-through the Origin Zone will be used. Note: this attribute is only relevant for Zones where "Is Pass-Through" is set to "Yes".
- Destination Zone ID: Numeric ID for the Destination Zone. This is the 'id' from the input shapefile, if provided by the user.
- Destination Zone Name: Name for the Destination Zone. This is the 'name' from in the input shapefile, if provided by the user.
- Destination Zone Is Pass-Through: "Yes" value indicates that only trips passing through the Destination Zone are represented in the frequency. "No" value indicates that only trips that end in the Destination Zone are represented in the frequency values.
- Destination Zone Direction: This refers to the direction in which trips pass-through the Destination Zone. Values are provided in degrees from 0 to 359, where 0 is due north, 90 is east, 180 is due south, etc. When creating Metrics, a range of -20/+20 degrees is applied to the Destination Zone Direction value. A value of "Null" refers to no direction filter and therefore all trips that pass-through the Destination Zone will be used. Note: this attribute is only relevant for Zones where "Is Pass-Through" is set to "Yes".
- Day Type: Average Day (average of traffic Monday through Sunday), Average Weekday (average of weekday traffic as defined by user), or Average Weekend Day (average of weekend traffic as defined by user).
- Day Part: Segments of the day defined by the user in intervals of hours to analyze traffic (All Day is always included as entire 24 hours). The Day Parts reflect the Origin Zones local time.
- Origin-Destination Traffic (frequency): Frequency value representing the volume of trips from the Origin Zone to the Destination Zone.

- Origin Zone Traffic (frequency): Frequency value representing all trips from the Origin Zone with no limitation on where they went.
- Destination Zone Traffic (frequency): Frequency value representing all trips to the Destination Zone with no limitation on where they came from.
- Avg Trip Duration (sec): Average time (in seconds) for the trips from the Origin Zone to the Destination Zone.

zone_frequencies_od_personal.csv & zone_frequencies_od_commercial.csv
===
These files contain information about each Zones used in the Project. The frequency represents all trips appropriate to each Zone.

- Vehicle Type: Type of vehicle analyzed with values of 'Personal' or 'Commercial'.
- Zone Type: Indicates if the Zone is an Origin or Destination Zone for this Project.
- Zone ID: Numeric ID for the Zone. This is the 'id' from the input shapefile, if provided by the user.
- Zone Name: Name for the Zone. This is the 'name' from in the input shapefile, if provided by the user.
- Zone Is Pass-Through: "Yes" value indicates that only trips passing through the Origin Zone are represented in the frequency. "No" value indicates that only trips that start or end in the Zone are represented in the frequency.
- Zone Direction: This refers to the direction in which trips pass-through the Zone. Values are provided in degrees from 0 to 359, where 0 is due north, 90 is east, 180 is due south, etc.
When creating Metrics, a range of -20/+20 degrees is applied to the Zone Direction value. A value of "Null" refers to no direction filter and therefore all trips that pass-through the Zone will be used.
Note: this attribute is only relevant for Zones where "Is Pass-Through" is set to "Yes".
- Day Type: Average Day (average of traffic Monday through Sunday), Average Weekday (average of weekday traffic as defined by user), or Average Weekend Day (average of weekend traffic as defined by user).
- Day Part: Segments of the day defined by the user in intervals of hours to analyze traffic (All Day is always included as entire 24 hours). The Day Parts reflect the Origin Zones local time.
- Zone Frequency: Frequency value representing all trips starting in, passing through, or ending in the Zone based on the Zone Type and the Zone Is Pass Through values.
If the Zone has an "Is Pass Through" value of yes, then the Zone Frequency is for all trips passing through the Zone. Otherwise, the Zone Frequency represents the trips starting in Origin Zones or ending in Destination Zones.

*_zone_set.(dbf|prj|shp|shx)
============================
These files comprise the shapefiles for the project's zone sets.

A shapefile consists of the following several files:
.shp file contains the feature geometries and can be viewed in a geographic information systems application such as QGIS.
.dbf file contains the attributes in dBase format and can be opened in Microsoft Excel.
.shx file contains the data index.
.prj file contains the projection information.

These shapefiles have the following attributes/columns:
- id: ID for the zone of interest as entered upon creation of zone set. This may be null as the field is optional.
- name: Name for the zone of interest.
- direction: Direction of travel in degrees where the trip passes through the zone. Values are from 0 to 360
  where 0 is north, 90 is east, and 225 is southwest.
- is_pass: 1 value indicates trips are expected to pass through the zone and that designation was used when running Metrics for this project. This indicates the Zone is likely a road segment, but not necessarily. 0 value indicates trips are expected to start or end in the zone and that designation was used when running Metrics for this project.
- geom: Polygon of the zone.

Notes
=====
OD Pairs with No Values
======================
If the frequency values for an OD pair for a specific time period (e.g. Average Weekday, Early Am) are below Vendor B's significance threshold, no results will be shown in the od_personal.csv & od_commercial.csv files.

Day Part Calculations
=====================
The Day Part calculations are done in relation to the Zones used in the analysis. The Origin-Destination Traffic values Day Parts are calculated in relation to the Origin Zone. The Day Part is determined by when Trips either Start in the Origin Zone or pass-by the centroid of the ORigin Zone, if the ORigin Zone is "Pass-Through" designated.
The Origin Zone Traffic values Day Parts are also calculated in relation to the Origin Zone, in similiar fashion. The Destination Zone Traffic value Day Parts are calculated in relation to the Destinaton Zone. The Day Part is determined by when Trips either end in the Destination Zone or pass-by the centroid of the Destination Zone, if the Destination Zone is "Pass-Through" designated.

Frequencies
===========
Frequency values represent trip activity but do not indicate actual number of trips or vehicles. The values are provided on an index. Personal and Commercial values use different indices. Projects in the US and Projects in Canada also use different indices.
For US Projects, a value of 500,000 on each index corresponds to average daily traffic on a stretch of Interstate 95 in the Mid-Atlantic.
For Canadian Projects, a value of 500,000 on each index corresponds to average daily traffic on a stretch of Highway 401 east of Toronto.

Comparing Frequencies
=====================
The frequency values for each vehicle type, weight class, and country are based on different sample populations and therefore cannot be compared with each other. Even though all of the Commercial weight classes use the same index, their frequency values cannot be compared with each other.

Device Ping Rate for Trips
==================
Projects with specified Device Ping Rates (values other than "All" in project.txt) filter out trips that have greater than the listed Device Ping Rate. This will reduce the the sample size used for the analysis. Frequency values for these projects use different indices and therefore cannot be compared to projects with Projects using "All" Device Ping Rates. As described in the "Frequencies" section above, Personal and Commercial values cannot be compared. Nor can Projects in the US and Projects in Canada.

Trip Type
=========
The project.txt specifies the type of Trips used in the analysis: Locked to Route Trips or Unlocked Trips. Unlocked Trips may not consistently align with roads depending upon the Device Ping Rate for Trips, the speed of the vehicle, and how curvy the roads are. Locked to Route Trips address this by aligning to the road segments of the most likely path taken for the set of points that comprise the Unlocked Trip.

7.7 ATRI – allencountyDataDictionary.txt

```
#################################################
##| Data Dictionay for allenCountySeqFull.csv |##
#################################################

-------------------
--| Description |--
-------------------

This file contains Global Positioning System (GPS) point data that has
been processed into sequence ordered by the truck's unique
identifier and the date/time stamp associated with each point. Each row
of this dataset essentially corresponds to two (2) GPS points. Variables
with a "to_" distinction corresponds to the next point in the sequence, ie.
as can be seen in the data the "to_readdate" value matches the succeeding
"readdate_from" value for each unique truck. During processing, if the next
point in the sequence corresponded to a different unique truck ID that
observation was removed so as not to have a "to_" value correspond to an
incorrect "_from" value for a truck indentifier.


----------------------------
--| Variable Definitions |--
----------------------------

truckid       : Unique truck identifier
readdate_from : Date/time stamp of GPS point
to_readdate   : Date/time stamp of succeeding GPS point
GEOID_from    : Census block identifier of GPS point
to_GEOID      : Census block identifier of succeeding GPS point
distance      : Great circle distance in miles between the two GPS points
secBtwPoints  : Time elapsed in seconds between the two GPS points
minBtwPoints  : Time elapsed in minutes between the two GPS points
hourBtwPoints : Time elapsed in hours between the two GPS points
speedBtwPoints: Space mean speed in miles per hour between the two GPS points
```

7.8 EE and EI/IE trip length distributions

The tables below provide EE and EI/IE trip length distributions for one minute bins. The values are expressed as percentages.

Table 35: EE trip length distributions - one minute bins

Time intervals	ODOT 2011	Vendor A:WTD	Vendor A:SP		ODOT 2011 Work/ Non-work	Vendor B Personal		ODOT 2011 Truck	Vendor C	Vendor B: Commercial
< 1	0.00%	0.00%	0.00%		0.00%	0.00%		0.00%	0.00%	0.00%
to 2	0.00%	0.00%	0.00%		0.00%	0.00%		0.00%	0.00%	0.00%
to 3	0.00%	0.00%	0.00%		0.00%	0.00%		0.00%	0.00%	0.00%
to 4	0.00%	0.00%	0.00%		0.00%	0.00%		0.00%	0.00%	0.00%
to 5	0.00%	0.00%	0.00%		0.00%	0.00%		0.00%	0.00%	0.00%
to 6	0.35%	11.06%	0.28%		0.40%	1.78%		0.28%	0.41%	4.28%
to 7	0.20%	8.05%	0.00%		0.33%	0.65%		0.00%	0.00%	0.75%
to 8	0.15%	26.04%	2.86%		0.23%	1.59%		0.02%	0.02%	2.80%
to 9	0.06%	5.63%	0.05%		0.11%	1.11%		0.00%	0.01%	0.17%
to 10	0.11%	3.23%	0.99%		0.17%	0.30%		0.04%	0.00%	0.12%
to 11	0.09%	7.49%	0.19%		0.09%	0.79%		0.08%	0.02%	0.16%
to 12	0.59%	7.44%	0.84%		0.54%	0.21%		0.68%	0.00%	0.10%
to 13	0.26%	4.15%	0.47%		0.41%	0.33%		0.05%	0.01%	0.08%
to 14	0.10%	6.32%	0.75%		0.16%	0.21%		0.01%	0.03%	0.03%
to 15	0.30%	0.72%	3.57%		0.47%	0.34%		0.06%	0.12%	0.18%
to 16	0.60%	3.53%	1.55%		0.50%	1.48%		0.74%	0.13%	0.69%
to 17	0.90%	0.62%	0.19%		1.29%	0.46%		0.33%	0.17%	0.42%
to 18	0.33%	0.83%	0.00%		0.53%	0.22%		0.04%	0.04%	0.05%
to 19	2.42%	0.68%	3.71%		3.39%	1.39%		1.01%	0.45%	1.02%
to 20	1.15%	0.65%	4.13%		1.33%	0.90%		0.89%	0.09%	0.58%
to 21	7.20%	3.51%	13.33%		6.94%	8.40%		7.57%	4.13%	6.00%
to 22	1.70%	0.62%	1.22%		2.04%	0.86%		1.20%	1.67%	2.20%
to 23	75.63%	5.23%	49.13%		71.63%	66.85%		81.42%	79.95%	69.63%
to 24	0.61%	0.84%	0.89%		0.64%	0.26%		0.58%	0.33%	0.50%
to 25	1.13%	1.19%	9.71%		1.24%	4.34%		0.98%	2.29%	4.67%
to 26	2.52%	0.81%	3.00%		3.00%	1.68%		1.82%	4.34%	2.89%
to 27	0.12%	0.23%	0.00%		0.16%	0.06%		0.06%	0.03%	0.08%
to 28	0.44%	0.32%	0.70%		0.58%	0.18%		0.23%	0.07%	0.31%
to 29	0.08%	0.14%	0.38%		0.11%	0.13%		0.04%	0.09%	0.09%
to 30	0.37%	0.31%	0.09%		0.43%	0.26%		0.27%	0.05%	0.12%
to 31	0.14%	0.11%	0.94%		0.20%	0.02%		0.04%	0.03%	0.05%
to 32	2.24%	0.06%	0.61%		2.84%	5.13%		1.38%	2.52%	1.90%
to 33	0.05%	0.03%	0.14%		0.06%	0.01%		0.04%	0.03%	0.01%
to 34	0.01%	0.03%	0.14%		0.02%	0.00%		0.00%	0.13%	0.03%
to 35	0.05%	0.03%	0.00%		0.06%	0.03%		0.03%	0.39%	0.07%
to 36	0.04%	0.02%	0.05%		0.05%	0.01%		0.03%	0.57%	0.02%

to 37	0.00%	0.03%	0.00%		0.00%	0.00%		0.00%	0.00%	0.00%
to 38	0.06%	0.03%	0.09%		0.04%	0.00%		0.08%	0.59%	0.01%
to 39	0.00%	0.00%	0.00%		0.00%	0.00%		0.00%	1.28%	0.00%
to 40	0.00%	0.01%	0.00%		0.00%	0.00%		0.00%	0.00%	0.00%
to 41	0.00%	0.00%	0.00%		0.00%	0.00%		0.00%	0.00%	0.00%
to 42	0.00%	0.00%	0.00%		0.00%	0.00%		0.00%	0.00%	0.00%

Table 36: EI/IE trip length distributions - one minute bins

Time intervals	ODOT 2011	Vendor A:WTD	Vendor A:SP		ODOT 2011 Work/ Non-work	Vendor B Personal		ODOT 2011 Truck	Vendor C	Vendor B: Commercial
< 1	2.07%	2.32%	3.92%		2.08%	4.63%		2.02%	4.92%	1.70%
to 2	3.96%	1.36%	3.56%		4.06%	2.42%		3.00%	5.32%	1.03%
to 3	10.72%	2.85%	2.90%		11.13%	5.09%		6.72%	14.53%	2.92%
to 4	7.11%	3.34%	2.40%		7.10%	3.73%		7.13%	7.06%	2.60%
to 5	3.50%	2.94%	2.19%		3.36%	3.02%		4.82%	3.84%	3.01%
to 6	5.16%	8.72%	9.40%		4.81%	5.25%		8.54%	24.90%	11.52%
to 7	2.27%	3.21%	3.34%		2.26%	3.33%		2.32%	2.50%	2.33%
to 8	4.22%	5.92%	6.83%		3.91%	9.33%		7.23%	4.75%	9.26%
to 9	4.21%	3.88%	3.26%		4.26%	5.07%		3.71%	2.05%	2.26%
to 10	4.16%	4.82%	4.29%		4.39%	4.29%		1.94%	1.07%	1.49%
to 11	4.65%	5.44%	4.83%		4.83%	4.93%		2.94%	2.22%	5.01%
to 12	4.98%	5.34%	5.21%		5.21%	5.56%		2.72%	5.55%	5.46%
to 13	3.69%	5.21%	5.33%		3.87%	5.05%		1.94%	2.44%	4.93%
to 14	4.13%	4.38%	4.72%		4.33%	3.39%		2.21%	0.98%	1.96%
to 15	4.09%	5.66%	5.30%		4.13%	4.79%		3.73%	1.80%	5.12%
to 16	4.37%	6.21%	6.53%		4.43%	3.21%		3.82%	1.41%	4.03%
to 17	4.04%	5.37%	5.15%		4.19%	3.97%		2.60%	0.77%	3.97%
to 18	3.90%	4.87%	4.37%		4.05%	6.59%		2.50%	6.50%	12.98%
to 19	2.72%	3.19%	2.81%		2.83%	2.29%		1.61%	1.51%	2.72%
to 20	2.87%	3.05%	2.68%		2.95%	2.77%		2.12%	3.68%	6.79%
to 21	2.58%	2.53%	2.18%		2.62%	1.83%		2.24%	0.57%	1.91%
to 22	2.71%	3.25%	3.35%		2.79%	3.01%		1.87%	0.37%	2.54%
to 23	1.67%	1.65%	1.30%		1.69%	2.90%		1.54%	0.72%	2.17%
to 24	1.52%	1.66%	1.50%		1.47%	0.88%		2.09%	0.16%	0.61%
to 25	1.34%	1.49%	1.43%		1.05%	0.91%		4.21%	0.10%	0.84%
to 26	0.76%	0.65%	0.54%		0.54%	0.77%		2.95%	0.14%	0.29%
to 27	0.37%	0.17%	0.28%		0.34%	0.42%		0.71%	0.03%	0.15%

to 28	0.45%	0.17%	0.16%		0.29%	0.19%		1.98%	0.02%	0.22%
to 29	0.43%	0.07%	0.06%		0.25%	0.09%		2.16%	0.03%	0.07%
to 30	0.28%	0.05%	0.03%		0.15%	0.04%		1.50%	0.00%	0.02%
to 31	0.16%	0.08%	0.05%		0.10%	0.01%		0.71%	0.01%	0.01%
to 32	0.22%	0.03%	0.03%		0.12%	0.00%		1.17%	0.01%	0.03%
to 33	0.21%	0.03%	0.03%		0.13%	0.08%		1.01%	0.01%	0.04%
to 34	0.18%	0.01%	0.02%		0.09%	0.06%		1.04%	0.00%	0.01%
to 35	0.15%	0.01%	0.01%		0.07%	0.10%		0.85%	0.00%	0.01%
to 36	0.07%	0.02%	0.00%		0.04%	0.00%		0.34%	0.00%	0.01%
to 37	0.03%	0.01%	0.02%		0.04%	0.00%		0.02%	0.00%	0.00%
to 38	0.02%	0.01%	0.00%		0.02%	0.00%		0.01%	0.00%	0.00%
to 39	0.01%	0.00%	0.00%		0.01%	0.00%		0.00%	0.00%	0.00%
to 40	0.01%	0.00%	0.00%		0.01%	0.00%		0.00%	0.00%	0.00%
to 41	0.00%	0.00%	0.00%		0.00%	0.00%		0.00%	0.00%	0.00%
to 42	0.00%	0.00%	0.00%		0.00%	0.00%		0.00%	0.00%	0.00%