Estimating Origin-Destination with Data Fusion: A Proof of Concept

OCTOBER 2017

U.S. Department of Transportation
Federal Highway Administration

TMIP
Better Methods. Better Outcomes.

Notice

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The U.S. Government assumes no liability for the use of the information contained in this document.

The U.S. Government does not endorse products or manufacturers. Trademarks or manufacturers' names appear in this report only because they are considered essential to the objective of the document.

Quality Assurance Statement

The Federal Highway Administration (FHWA) provides high-quality information to serve Government, industry, and the public in a manner that promotes public understanding. Standards and policies are used to ensure and maximize the quality, objectivity, utility, and integrity of its information. The FHWA periodically reviews quality issues and adjusts its programs and processes to ensure continuous quality improvement.

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle	5. Report Date
Estimating Origin-Destination Using Data Fusion: A Proof of Concept	October 2017
	6. Performing Organization Code

7. Authors	8. Performing Organization Report No.
Vince Bernardin, Jr., PhD Hadi Sadrsadat, PhD John Gliebe, PhD Nagendra Dhakar, PhD Steven Trevino	

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
RSG 55 Railroad Row White River Junction, VT 05001	**11. Contract or Grant No.** DTFH61-12-D-00013, T-12002

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
United States Department of Transportation Federal Highway Administration 1200 New Jersey Ave. SE Washington, DC 20590	March 2014- October 2017
	14. Sponsoring Agency Code HEPP-30

15. Supplementary Notes

The project was managed by Task Manager for Federal Highway Administration, Sarah Sun, who provided detailed technical directions.

16. Abstract

This fourth volume of Bridging Data Gaps: A TMIP Series on Understanding Origin-Destination Data presents methods for developing data driven estimates of travel demand patterns for traffic forecasting. These methods take advantage of both traditional traffic counts as well as new passively collected Big Data on origin-destination patterns. The volume provides detailed methods for data validation and quality assurance, including methods for checking the reasonableness and consistency of traffic counts on a highway network and methods for expanding passively collected OD data. It is critically important that passive OD data be properly expanded to represent the travel of interest because passively collected data does not constitute a random sample and is not generally representative. In particular, growing evidence indicates the systematic over-representation of longer trips and activities relative to shorter ones in passive OD data. This volume presents several methods for correcting this systematic bias in passive OD data using traffic count data as well as methods for using this data in traffic modeling and forecasting.

17. Key Words	18. Distribution Statement
Origin-destination data; origin-destination estimation, travel surveys, household surveys, trip diary surveys, transit surveys, Big Data for transportation, data fusion	No restrictions.

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	78	N/A

SI* (MODERN METRIC) CONVERSION FACTORS

APPROXIMATE CONVERSIONS TO SI UNITS

Symbol	When You Know	Multiply By	To Find	Symbol
		LENGTH		
in	inches	25.4	millimeters	mm
ft	feet	0.305	meters	m
yd	yards	0.914	meters	m
mi	miles	1.61	kilometers	km
		AREA		
in^2	square inches	645.2	square millimeters	mm^2
ft^2	square feet	0.093	square meters	m^2
yd^2	square yard	0.836	square meters	m^2
ac	acres	0.405	hectares	ha
mi^2	square miles	2.59	square kilometers	km^2
		VOLUME		
fl oz	fluid ounces	29.57	milliliters	mL
gal	gallons	3.785	liters	L
ft^3	cubic feet	0.028	cubic meters	m^3
yd^3	cubic yards	0.765	cubic meters	m^3
		NOTE: volumes greater than 1000 L shall be shown in m^3		
		MASS		
oz	ounces	28.35	grams	g
lb	pounds	0.454	kilograms	kg
T	short tons (2000 lb)	0.907	megagrams (or "metric ton")	Mg (or "t")
		TEMPERATURE (exact degrees)		
°F	Fahrenheit	5 (F-32)/9 or (F-32)/1.8	Celsius	°C
		ILLUMINATION		
fc	foot-candles	10.76	lux	lx
fl	foot-Lamberts	3.426	$candela/m^2$	cd/m^2
		FORCE and PRESSURE or STRESS		
lbf	poundforce	4.45	newtons	N
lbf/in^2	poundforce per square inch	6.89	kilopascals	kPa

APPROXIMATE CONVERSIONS FROM SI UNITS

Symbol	When You Know	Multiply By	To Find	Symbol
		LENGTH		
mm	millimeters	0.039	inches	in
m	meters	3.28	feet	ft
m	meters	1.09	yards	yd
km	kilometers	0.621	miles	mi
		AREA		
mm^2	square millimeters	0.0016	square inches	in^2
m^2	square meters	10.764	square feet	ft^2
m^2	square meters	1.195	square yards	yd^2
ha	hectares	2.47	acres	ac
km^2	square kilometers	0.386	square miles	mi^2
		VOLUME		
mL	milliliters	0.034	fluid ounces	fl oz
L	liters	0.264	gallons	gal
m^3	cubic meters	35.314	cubic feet	ft^3
m^3	cubic meters	1.307	cubic yards	yd^3
		MASS		
g	grams	0.035	ounces	oz
kg	kilograms	2.202	pounds	lb
Mg (or "t")	megagrams (or "metric ton")	1.103	short tons (2000 lb)	T
		TEMPERATURE (exact degrees)		
°C	Celsius	1.8C+32	Fahrenheit	°F
		ILLUMINATION		
lx	lux	0.0929	foot-candles	fc
cd/m^2	$candela/m^2$	0.2919	foot-Lamberts	fl
		FORCE and PRESSURE or STRESS		
N	newtons	0.225	poundforce	lbf
kPa	kilopascals	0.145	poundforce per square inch	lbf/in^2

*SI is the symbol for the International System of Units. Appropriate rounding should be made to comply with Section 4 of ASTM E380.
(Revised March 2003)

Estimating Origin-Destination Using Data Fusion: A Proof of Concept

Original: April 2017
Final: October 2017

Prepared for:

Federal Highway Administration

Table of Contents

List of Figures

List of Tables

List of Abbreviations and Symbols

Abbreviations

AADT	Annual Average Daily Traffic
AB	Activity-Based
ATRI	American Transportation Research Institute
CV	Coefficient of Variation
FAST	Fixing America's Surface Transportation (Act)
FHWA	Federal Highway Administration
GPS	Global Positioning System
LBS	Location-Based Services
MPO	Metropolitan Planning Organizations
NCHRP	National Cooperative Highway Research Program
OD	Origin-Destination
ODME	Origin-Destination Matrix Estimation
RMSE	Root Mean Squared Error
STOPS	Simplified Trips-on-Project Software
TDOT	Tennessee Department of Transportation
TMIP	Travel Model Improvement Program
TSTM	Tennessee Statewide Travel Model
ZOV	Zero-Occupant Vehicle

1.0 Introduction

1.1 Disclaimer

The views expressed in this document do not represent the opinions of FHWA and do not constitute an endorsement, recommendation or specification by FHWA. The document is based solely on the research conducted by RSG.

1.2 Acknowledgments

This volume is a collaboration between transportation professionals at FHWA, FTA, the Tennessee Department of Transportation, and RSG.

1.3 Introduction and Overview

Despite many years of effort improving the practice of travel forecast modeling, many forecasting models still struggle to accurately represent current year travel patterns. The most critical deficiency occurs with the representation of trip origin-destination (OD) patterns. This critical difficulty in replicating the spatial distribution of trips is widely acknowledged in both practice and research as the largest source of error in travel forecasting. Moreover, an accurate understanding and representation of OD patterns is critical to many important analyses such as whether travelers might change modes, pay a toll to use an express lane, or send their autonomous vehicle home rather than pay to park it.

The preceding three volumes of Bridging Data Gaps: A TMIP Series on Understanding Origin-Destination Data review both traditional and emerging sources of information on OD patterns. An emphasis is placed on understanding the limitations of the various types of data to help encourage thoughtful and appropriate use of these data while also highlighting the promise of new datasets and methods. This fourth volume serves as a complement to the preceding volumes and demonstrates how various sources of data are already being used, particularly in combination, to produce more data-driven forecasts to support planning, design, and operational analysis.

This volume presents a broad set of methods for processing and combining traffic count and large-scale passive OD data and using these with travel demand models to produce data-driven traffic forecasts. This topic has evolved rapidly over recent years and will likely continue to do so for some time. That said, the information presented in this volume should remain a valuable resource for understanding practical methods and how they can be used to produce robust, data-driven highway forecasts. In turn, these forecasts may be used to provide accurate and relevant information to help answer critical questions about how transportation is changing or may change in the future.

The methods explored here take advantage of both traditional traffic counts and new passively collected Big Data on OD patterns. This volume provides detailed methods for data validation and quality assurance, including methods for checking the reasonableness and consistency of traffic counts on a highway network and methods for expanding passively collected OD data. Data from the development of the Tennessee Statewide Travel Model (TSTM) is used to illustrate many important considerations and methods, and to provide in-depth examples of practical techniques to support data-driven traffic forecasting. This serves as a proof of concept to demonstrate how

traffic counts and passive OD data can be reconciled and combined to produce estimates of current OD patterns for both automobiles and trucks.

As discussed in Volume 2 and Volume 3 of this series, expanding the sample data to represent the full population of interest continues to be critically important when conducting data fusion. Passively collected data does not constitute a random sample and is not generally representative. Growing evidence indicates the systematic overrepresentation of longer trips and activities relative to shorter ones in passive OD data. This volume presents several methods for correcting this systematic bias in passive OD data using traffic count data and methods for using this data in traffic modeling and forecasting.

1.4 Data-Driven Traffic Forecasting Methods

Replicating trip origin-destination patterns widely acknowledged in both practice[1] and research (1) as the largest source of error in travel forecasting. Although recent destination choice models offer some important improvements over traditional gravity models, destination choice models still struggle to reproduce observed travel patterns. Data-driven forecasting methods, such as those presented in this volume, can be used with models as discussed in Section 5.0; however, these methods rely more on direct data than on models for forecasting travel patterns, particularly in the challenging spatial dimension.

The desire to ground traffic forecasts in data is not new. Many of the elements and general approaches presented in this volume—such as forecasting by pivoting off traffic counts and using counts to update OD matrices—have been used for many years and recommended as best practices for traffic forecasting in *National Cooperative Highway Research Program (NCHRP) Report 255: Highway Traffic Data for Urbanized Area Project Planning and Design* and, more recently, in its update, *NCHRP Report 765: Analytical Travel Forecasting Approaches for Project-Level Planning and Design*. (2)

While the NCHRP 255 and NCHRP 765 reports focus on project-level forecasting, similar data-driven methods have also been incorporated in general forecasting models. Data-driven models have become a common practice in statewide modeling (e.g., Michigan, Indiana, Tennessee, Florida); these models pivot off base-year OD matrices that have been improved by count information given the even greater difficulty in representing intercity OD patterns with gravity and destination choice models than representing local patterns in urban models.

Data-driven modeling is also widespread for metropolitan modeling outside of the United States; in fact, this practice is required in the United Kingdom.[2] Awareness of this in the United States has increased recently due to greater global interaction and communication, and through TMIP webinars by RAND Europe showcasing their work in both Europe and Australia.

At an urban level in the United States, the Federal Transit Administration's Simplified Trips-on-Project Software (STOPS) has demonstrated the ability of data-driven methods to rapidly prepare plausible estimates of transit ridership, replacing the traditional, time-consuming process of model

[1] *Spatial interaction models*, Travel Forecasting Resource, Spatial interaction models link.

[2] Transport Analysis Guidance, UK Department for Transport, Transport analysis guidance: WebTAG link.

development, validation, and application. STOPS uses four types of data to generate estimates of project ridership:

1. Origin-destination total person trip flows from the Census Transportation Planning Package
2. Origin-destination transit trip flows from transit on-board surveys
3. Level-of-service measurements from transit schedules
4. Current year transit counts to represent observed transit usage of individual stations and routes

STOPS runs in an iterative manner to generate an initial estimate of ridership and then adjusts these ridership estimates to match counted ridership at whatever level of detail is available in the count database. Once this model is calibrated, it is then applied to future year scenarios with and without the proposed transit improvement.

Finally, data-driven forecasting methods are now being applied in metropolitan modeling in the United States. Metropolitan planning organizations (MPOs) and state departments of transportation have begun to capitalize on the commercial availability of large-scale, aggregated, anonymous, passively collected OD data, also commonly referred to as "Big Data," and referred to as passive OD data throughout this volume. (3,4,5,6,7,8,9,10,11,12,13,14,15) The performance of new models of this type using these data is promising and is the focus of this volume.

1.5 New OD Data Sources to Support Data-Driven Highway Forecasting

Unlike traditional travel demand models, which rely almost exclusively on survey data, or earlier data-driven forecasting methods, which often relied primarily or exclusively on traffic counts, contemporary data-driven traffic forecasting incorporates passive OD data. Passive OD data is large in scale (generally including millions of trips), typically aggregated (to protect privacy concerns and for data manageability), anonymous (not including any traveler characteristics), and passively collected. Cost is one of the main advantages of these passive datasets, which can provide OD data more cost effectively than surveys. However, the completeness or adequacy of the sample in the spatial dimension provides another powerful motivation for the use of these datasets. While traditional travel surveys typically provide observations on 2% or fewer of the OD pairs in a region, new sources of Big Data can provide observations covering more than one-quarter or one-third of all OD pairs. This order of magnitude difference in the completeness of the data provides a more complete picture of spatial travel patterns in a region that traditional data could not.

The two currently predominant Big Data sources are cellular communications data and Global Positioning System (GPS) data, although Bluetooth is also a source of such data for limited applications (i.e., freeway corridors, external cordons). Recently, in early 2017, Location-Based Services (LBS) data has also become available. Each technology requires its own equipment and has its own limitations. For instance, cellular phone tower triangulation has limited resolution based on the spacing of towers and relies on communications between devices and towers that is not optimized for transportation data needs. GPS devices can provide accurate locational data, but sometimes ID persistence is an issue that can limit data processing techniques. Bluetooth

transceivers are required to detect Bluetooth-enabled devices and must be deployed on site to collect the data. Despite these limitations, these new technologies provide information on millions of trips, which produces rich trip tables for travel forecasting purposes. (Big Data can provide 100 to 1,000 times as many trips as observed in a survey.)

AirSage, StreetLight, and American Transportation Research Institute (ATRI) are examples of passive OD data providers. AirSage currently provides data only on total traffic. ATRI provides data only for heavy trucks. And StreetLight provides data broken out by cars and trucks. By using these new data sources, data-driven traffic forecasting methods can avoid relying on the unreliable trip distribution components of travel demand models and produce better assignment results and forecasts, especially when used in combination with traffic counts. At the same time, passive OD data has limitations, including potential biases that must be addressed and corrected to arrive at robust and reliable forecasts.

Section 3.0 of this volume will further explore these new passive OD data sources and provide detailed information necessary for processing, cleaning, and combining these data for data-driven forecasting.

1.6 The Necessity of Traffic Counts to Expand and Validate Big Data

Contemporary data-driven forecasting no longer relies entirely on traffic counts, but traffic counts remain the best data source for validating OD flows. Ultimately, highway volumes are the desired result of a traffic forecasting exercise, and the forecasting model or procedure should be verified against ground counts for the current or base year. More accurate counts and more complete coverage provide greater confidence in the forecasting process.

Traffic counts remain indispensable to data-driven highway forecasting due to the need to expand passive OD data. New technologies can provide very large data samples, but the data remain samples and must be expanded to correct for any biases in the sample data. Methods for expanding passive OD data using only estimated sample penetration rates (as opposed to counts) have been proposed and are sometimes employed by data vendors. However, these methods ignore many possible biases, particularly location-specific biases unrelated to sample penetration (e.g., poor coverage or line-of-sight in an area) and biases related to trip distance or duration (e.g., data dependent on signaling events is more likely to observe longer trips and less likely to observe shorter trips). As a result, penetration-based methods often produce only mediocre agreement with traffic counts. Traffic counts provide the best (and generally unbiased) data on the actual total amount of traffic on the road in various locations. Therefore, methods that leverage count data for expanding passive OD data can produce far better results.

Traffic count data must be cleaned and verified given its importance in a data-driven forecasting process. Section 2.0 of this volume will focus on this easily overlooked—but critical—task of validating traffic counts, which includes reviewing methods for checking the logical consistency of counts with other roadway attributes, reviewing the temporal consistency of counts at the same station from multiple years, and reviewing the spatial consistency of counts on adjacent and nearby roadway segments.

1.7 Combining Network and Demand Data

Section 4.0 of this volume focuses on the reconciliation and combination of passive OD data and traffic count data to produce expanded OD trip tables for traffic assignment. Several approaches are discussed and illustrated, including both origin-destination matrix estimation (ODME) and non-ODME-based methods.

It is important to recognize and emphasize that ODME can be misused. This volume illustrates how to properly use ODME and avoid instances of misuse. This volume also calls attention to the fact that the seed OD matrix for ODME is—in some ways—more important than the traffic counts. Responsible and proper use of ODME always considers the fit of the ultimate solution against both the traffic counts and the original seed. Ultimately, Section 4.0 illustrates how ODME and other methods can be used responsibly to reconcile and combine passive OD data and traffic counts to produce OD trip tables better grounded in data.

1.8 Case Study Introduction: Tennessee Statewide Travel Model

This volume enables forecasting professionals to apply the technical methods and techniques presented here to their own data and forecasting needs. To facilitate this goal, this volume includes examples of how these principles and techniques have been applied. For this purpose, this volume uses several datasets, but especially data from the Tennessee Department of Transportation (TDOT). This volume tracks TDOT's data development process used to support their new statewide model.

The TSTM was originally developed in 2004 to support TDOT's previous long-range plan update. Since then, the TSTM has also been successfully used to support several studies throughout the state. However, given its limited resolution—it did not include all state routes in its network—it became apparent that improvements were necessary to fully leverage the model for project-level planning and programming. Moreover, TDOT also recognized the opportunity for the TSTM to help generate performance measures for use in decision assistance and to support the continuous performance-based planning and programming process mandated by MAP-21 and the FAST Act.

Therefore, TDOT undertook an effort to update the TSTM to support their new long-range plan update, to provide greater support for project-level forecasting and regional planning for rural areas outside MPO areas, and to generate peak-hour volumes and other performance measures to support a performance-based planning and programming process in accordance with current transportation legislation.

The TSTM is one of the first applications of passive OD data derived from cellular phone data to support the development of a statewide travel model. The TSTM also includes a truck model supported by the purchase of an eight-week dataset of truck GPS-based OD data from ATRI. The data included information from over 234,000 individual trucks on over 6.5 million truck trips representing approximately 11% of the trucks on the road for 56 days.

The data-driven approach used in the TSTM's development led to good model performance. The model's highway assignment achieved impressive validation statistics versus traffic counts for a statewide model, including a 37% root mean squared error (RMSE).

2.0 Traffic Counts

Traffic counts are best source of information on the total amount of traffic on a road, but traffic counts are not infallible. Traffic count data associated with travel model networks can include error from several sources. Sample error results when 24- or 48-hour counts are used as estimates of average daily traffic. Since traffic varies daily, individual counts may be higher or lower than the true average traffic on the roadway segment. Mechanical errors can also result from the various counter devices/technologies employed to take the count. Data processing errors can also result in traffic count data being associated with the wrong segment in the roadway network or a bidirectional count being associated with one segment of a dualized one-way pair (or vice versa). When counts are adjusted for seasonality and day of the week, these calculations are also subject to potential error.

It is important to verify traffic count data to identify and minimize errors to the extent possible before using count data to expand passive OD data. Neglecting this data validation task risks combining errors from the traffic count data with errors in the Big Data rather than using the relatively accuracy of count data to improve and reduce errors in passive data.

While many manual and automated checks can be performed on traffic count data, these procedures can be grouped into three types of checks on the consistency of the traffic count data with other information and itself. First, checks of the logical consistency of counts with other roadway attributes provide validation of the external validity of the count data. Second, checks of the internal validity of the traffic count data with itself can be divided between checks on the temporal consistency of counts at the same station from multiple years, and checks on the spatial consistency of counts on adjacent and nearby roadway segments. Each of these types of checks are explained in additional detail in the following sections.

2.1 Logical Consistency of Counts with Other Roadway Attributes

TDOT developed automated methods for identifying inconsistencies in a highway network count database using a count consistency checking tool. The tool was programmed into TDOT's existing travel demand modeling platform using TransCAD and GISDK scripting.

The tool is set up to identify potential problems with counts in the network and create a list of links for the user to review. For each of the checks described below, the user may specify error tolerances or thresholds to reflect the amount of acceptable error. The tool will generate a list of suspected problem links (errors) and these locations can be mapped. It is left to the user to review the list to determine whether the suspected problem counts really are errors and, if so, identify the source of the problem and a potential solution.

The first check is to identify counts that may be considered outliers when compared to coded link capacities. While daily counts that approach or even exceed the coded daily capacity for a link are clear outliers, there may be cases in which counts appear to be too low relative to daily capacity and should be reviewed to make sure it is properly coded. Because link capacities are typically coded based on facility and area type designations and number of lanes, this comparison implicitly takes these factors into consideration when determining whether a count is plausible. Accordingly, if an inconsistency is identified, the solution could point to either the count or the capacity coding designation as the source of the discrepancy.

The automated check compares link counts against corresponding link capacities. The comparison is performed by direction and only for the links with original counts. (The method described below discusses the procedure of count propagation, which enables more complex spatial consistency checks.)

First, for each link direction (AB or BA), count and capacity values are obtained. The obtained count is compared against low- and high-capacity thresholds as below:

Threshold low * capacity <= count <= Threshold high * capacity

where “threshold low” and “threshold high” are user inputs. Threshold values of 0 and 1, respectively, might be considered conservative default values.

The count is considered reasonable if it falls within the range. The check is performed by direction and different message codes (“msg”), as described in Table 1, are reported in an output file for user’s review. The output also contains count-to-capacity ratio by link direction. The output contains only those links that have original annual average daily traffic (AADT) counts.

Table 1: Link Capacity-based Checks Output

Msg	Label
0	No count
1	Count is reasonable
2	Count is low
3	Count is high
4	Count/capacity is not available
5	Count is on unexpected direction

2.2 Internal Temporal Consistency of Counts

The ability to check the internal temporal consistency of count data requires count data over time. In the past, this was often an obstacle to the use of this method as it was common for only a single daily count to be available for any given roadway segment or count station. However, it is increasingly common to have either hourly count data or multiple years of count data for a location, or both. If such information is available, it can be used to check the consistency of counts at a location over time.

Hourly count data can help identify both mechanical and data processing errors. Clearly irrational variations (e.g., a large spike in volume at 3:00 a.m. or extremely large hourly changes in volume throughout the day) in hourly count data can indicate mechanical errors. Such irregularities in hourly volumes are valuable because they can help identify mechanical errors in cases when the total daily volume may still have been plausible (but potentially wrong). However, these irregularities must be large for there to be confidence that they are truly erroneous and not legitimate hourly variations in traffic. Irrational patterns can also indicate data processing errors. For instance, if a one-way link from suburban residential areas toward the central business district

shows high volumes in the afternoon and low volumes in the morning (when the reverse would be expected) it indicates a likely processing error in which directional volumes were associated with the wrong directional link in the roadway network. However, detecting these types of errors can be somewhat challenging since they can require information on network topology.

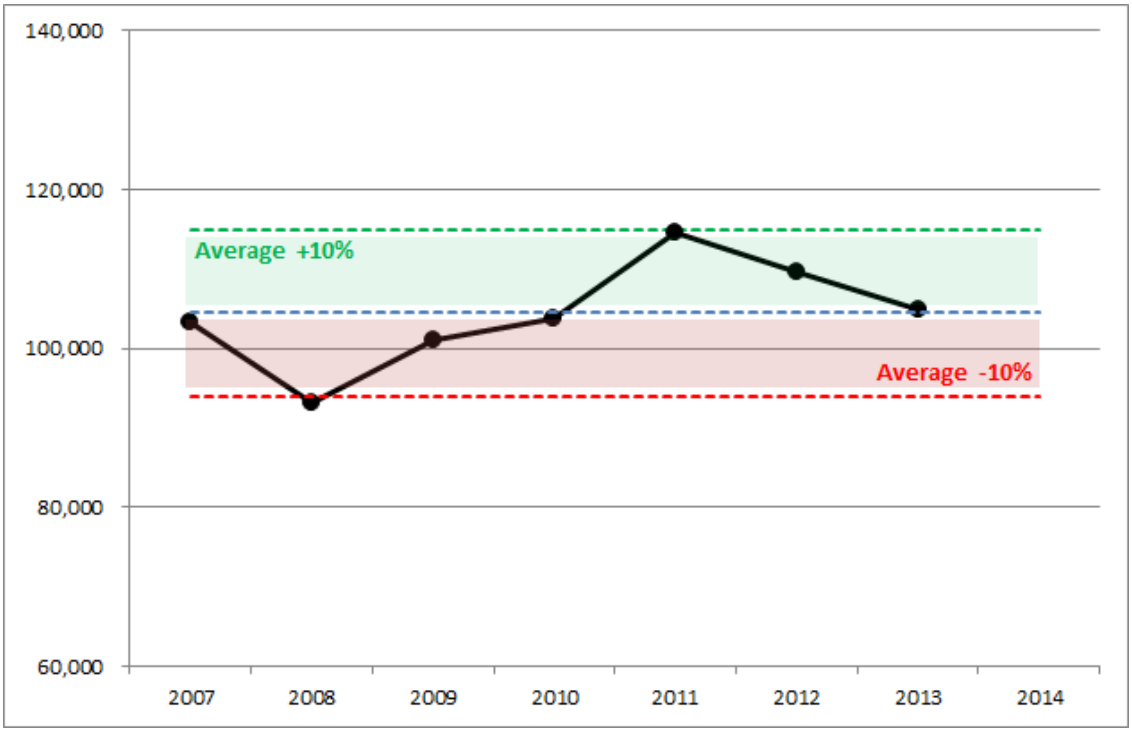
Count data over several years is much more valuable for verifying the internal validity of count data since year-to-year variability in counts is less than hour-to-hour variability and patterns are simple and not cyclic. Careful analysis of multiyear count data can help identify errors of any source and distinguish likely errors them from more plausible changes owing to general trends toward higher (or lower) traffic on a roadway. This approach was used successfully in the development of the TSTM.

TDOT supplied a count database for the State of Tennessee containing 12,297 count stations with traffic counts from 1983 through 2013. The analysis focused on the five-year period from 2008 to 2012 focused on the model's 2010 base year. The data validation involved a two-step process. First, for each station, a front-weighted mean AADT was calculated. Counts from 2012 were weighted 5, and each earlier year weighted one less, so 2008 was weighted 1. Then, for each station, each year's count was compared to the front-weighted mean for that station and retained or thrown out using the following criteria based on the Federal Highway Administration's (FHWA's) calibration standards:

Volume < 1,000	- acceptable difference =	+/- 200%
Volume < 2,500	- acceptable difference =	+/- 100%
Volume < 5,000	- acceptable difference =	+/- 50%
Volume < 10,000	- acceptable difference =	+/- 25%
Volume < 25,000	- acceptable difference =	+/- 20%
Volume < 50,000	- acceptable difference =	+/- 15%
Volume > 50,000	- acceptable difference =	+/- 10%

A front-weighted mean was used to allow for legitimate trends (basic steady increases or decreases) in traffic, so that if years were thrown out for stations with significant growth, the years thrown out would be the earlier, lower year and not the more recent, higher ones.

Figure 1: Example of I-40 Count Station Volumes in Davidson County



After this initial step, the (un-weighted) mean count volume, standard deviation, and coefficient of variation (CV) were calculated for each station with good years. For stations with only 2012 counts (unavailable throughout 2008–2011), 2013 count volumes were also used to calculate the CV. Stations were dropped as unreliable/highly variable/questionable if the CV was greater than 15% and the standard deviation was greater than 100. This composite criterion ensured that counts were only filtered out if their variation was substantial in both relative and absolute terms. Counts on low-volume facilities—in particular, those that might have higher coefficients of variation but still only vary within a range of 50 or so vehicles per day—were included as valid data. Of the over 12,000 count stations, 213 (approximately 1.7%) of the stations were removed due to this process. Inspection of a subset of the stations that were removed indicated several issues, including facilities that were affected by major reconstruction projects/detours and some stations with inexplicable but clearly unlikely variation in volumes.

2.3 Internal Spatial Consistency of Counts

There are multiple methods for checking the spatial consistency of counts; the most common (by far) is visual inspection. Sometimes simple visualization is used in a GIS context by which the color or width of a roadway segment is based on the size of its traffic count and an analyst attempts to canvas the model network in a systematic way to detect potentially inconsistent counts. More frequently, the spatial consistency of counts is only checked on a case-by-case basis after comparison of model volumes to counts reveals bad agreement between the model and counts in one location but good agreement between the model and counts nearby. While it is clearly important to check counts in this instance, it is also clearly preferable to identify the problem earlier and without relying too much on modeled volumes. For this reason, a tool to check the spatial consistency of counts was developed for Tennessee. The remainder of this section

documents the method used by this tool. The tool approaches the general problem of checking the spatial consistency of counts by using conservation-of-flow requirements at intersections; in order to do this, special consideration must be given to the coverage and propagation of counts on the network and locations where count data are missing.

2.3.1 Count Propagation

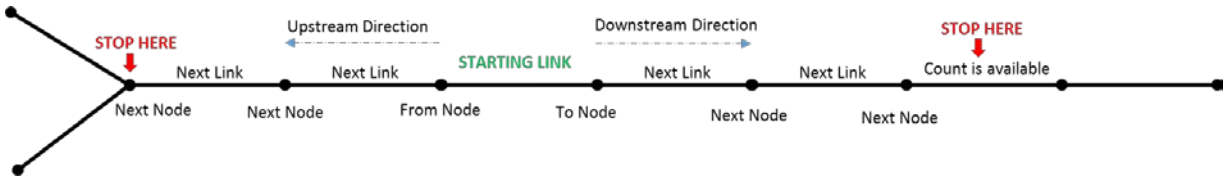
In a roadway network, a road segment may be represented by multiple links. However, the count collected on the segment is generally coded on one link. The remaining links do not get a count assigned to them. This may provide a false impression of having comparatively low count coverage on the network. In addition, this could be problematic while performing consistency checks at an intersection where a roadway link (an approach) may not have a count assigned to it even though a count is available for the approach (assigned to a different upstream link representing the same roadway segment).

The process of assigning (propagating) counts from a coded link to the other links of the roadway segment is referred to as count propagation. After count propagation, all links representing the segment have the same count assigned to them.

This section discusses the count propagation methodology employed in the Tennessee tool. Different measures to describe count coverage are provided in the subsequent section, which also includes count coverage statistics for the Tennessee statewide model network. This is followed by a more detailed section covering conservation-of-flow methods for determining spatial consistency.

The tool traverses the network and propagates counts from their actual coded locations to other adjacent locations. A graphical representation of the count propagation methodology implemented in the tool is shown in Figure 2.

Figure 2: Count Propagation Methodology



At a higher level, the tool loops through every link in the analysis region. Each of those links is called a “starting link.” For a starting link, the tool first finds its “From” and “To” nodes. The two nodes are then used as starting points to traverse in backward (upstream) and forward (downstream) directions. Each direction is traversed in the same manner and as described below.

For a node (one direction), a “link set” consisting of links connected to the node is obtained. Centroid connectors, if any, are removed from the “link set.” If the resulting “link set” consists of only two links (including the starting link), then a “next link” is determined by removing the starting link from the set (by comparing link IDs). A “next link” is the link at the node that is not the current link. End nodes of the “next link” are then used to find a “next node.” Like the “next link,” a “next node” is the node of the “next link” that is not the current node. Again, a “link set” is obtained by finding links connected to the “next node” and a “next link” is determined.

This process of finding a "next link" and a "next node" continues until one of the following is encountered:

- "Next link" with a count.
- "Next node" with a link set consisting more than two links.

After the network is traversed in both directions (upstream and downstream), a check is performed to see if any traversal direction encountered a count. If yes, then the traversed links (tracked during the process) are assigned with the count. In a scenario where both directions encounter counts, the two counts are compared. If the counts are not equal, they are reported as "conflicting counts" in an output report. Otherwise, the counts are assigned to the traversed links.

As traversed links are checked for counts (and assigned if available), they are not considered as starting links. Only links with no counts are processed to find counts.

The tool outputs a CSV file ("LinksWithPropagatedCounts.csv") consisting of roadway links with respective propagated counts (if any). The new counts are also attached to the roadway network database in two new count fields (by direction). The new fields are named as "_new" and attached to the original count field names. In addition, a field "msg" is also added to provide user with more information about the type of count. Table 2 presents the values used in the field and respective labels.

Table 2: Count Propagation Message Description

Value	Label
0	No count
1	Existing coded count
2	Propagated count
3	Conflicting counts

Coverage Statistics

The tool generates several statistics to describe count coverage in the network. The following measures are reported by the tool:

1. Before and after count propagation network level summary (number of links, lane miles, average count, average daily capacity) by:
    a. Count type (before and after count propagation).
    b. Link type (with count or without count).
2. After count propagation summary by:
    a. Functional class.
    b. Link type (with count or without count).
3. Summary of interchanges by:
    a. Interchange type (3-way, 4-way, and 5-way).
    b. Count availability (counts on all approaches, missing on approach counts, missing more than one approach count).

4. ODME summary—average number of shortest paths on links with counts and without counts. A statistic useful for determining whether counts are located on links that tend to provide the most connectivity between origins and destinations, hence useful for ODME. The number of shortest paths was based on an assignment of a trip table of "1s" to the free-flow statewide network.

The coverage statistics generated for the Tennessee statewide model network are shown in Table 3, Table 4, Table 5, Table 6, Table 7, and Table 8.

Table 3: Tennessee Network Summary (Existing Counts)

LinkType	NumLinks	(%)	LaneMiles	(%)	AADT_AB*	AADT_BA*	Cap_AB*	Cap_BA*
With Counts	12,830	65.49%	2,658.36	76.22%	3,152.74	4,588.72	24,500.50	20,610.48
Without Counts	104,891	34.51%	30,289.84	23.78%	0.00	952.00	25,041.21	20,508.15
TOTAL	117,721	100.00%	32,948.20	100.00%				

*These are averages across all links for a link type.

Table 4: Tennessee Network Summary (Propagated Counts)

LinkType	NumLinks	(%)	LaneMiles	(%)	AADT_AB*	AADT_BA*	Cap_AB*	Cap_BA*
With Counts	77,097	65.49%	25,113.96	76.22%	2,815.59	4,147.70	23,306.73	20,002.47
Without Counts	40,624	34.51%	7,834.24	23.78%	0.00	0.00	28,162.17	21,500.15
TOTAL	117,721	100.00%	32,948.20	100.00%				

*These are averages across all links for a link type.

Table 5: Tennessee Count Propagation Summary by Link Type and Functional Class

FUNCCLASS	NumLinks	(%)	LaneMiles	(%)	AADT_AB*	AADT_BA*	Cap_AB*	Cap_BA*
1: Rural Principal Arterial—Interstate	2,362	78.65%	1,252.34	90.60%	18,465.63	16,829.20	53,466.23	44,889.23
2: Rural Principal Arterial—Other	5,154	67.42%	1,823.20	69.49%	4,106.39	4,241.36	37,350.84	26,009.62
6: Rural Minor Arterial	7,113	72.64%	2,543.03	76.60%	2,354.94	2,381.50	18,351.37	16,824.81
7: Rural Major Collector	10,736	80.84%	4,369.09	84.33%	985.19	988.26	15,703.39	15,536.34
8: Rural Minor Collector	20,213	87.89%	9,396.85	88.59%	417.47	418.38	15,634.53	15,617.88
9: Rural Local	289	74.10%	134.29	78.57%	635.34	635.34	15,685.61	15,791.05
10: Unknown	11	0.64%	1.94	0.76%	1,703.14	1,992.33	11,216.31	6,479.87
11: Urban Principal Arterial—Interstate	2,525	54.49%	577.92	68.12%	41,089.37	38,714.37	69,510.96	57,408.74
12: Urban Principal Arterial—Other Freeways & Expressways	929	50.49%	190.02	61.39%	16,330.54	19,865.86	56,375.33	41,913.59
14: Urban Principal Arterial—Other	7,731	50.87%	1,262.00	56.60%	9,575.67	9,558.75	34,988.62	24,171.94
16: Urban Minor Arterial	10,154	55.42%	1,684.07	61.26%	4,429.45	4,531.83	20,727.35	17,934.51
17: Urban Collector	9,390	68.93%	1,783.77	74.15%	1,698.71	1,705.06	15,155.38	14,446.36
19: Urban Local	456	74.15%	93.58	79.99%	1,479.78	1,479.99	14,912.52	14,607.55
20: Unknown	18	0.41%	1.47	0.27%	10,991.22	5,252.50	12,128.39	2,400.65
91: 1-Lane Roundabout	16	27.59%	0.39	33.44%	4,087.58	2,748.88	6,594.76	1,637.11
TOTAL	**77,097**	**65.49%**	**25,113.96**	**76.22%**				

*These are averages across all links for a functional class.

Table 6: Tennessee Count Propagation Summary of Links with No Count

FUNCCLASS	NumLinks	(%)	LaneMiles	(%)	AADT_AB*	AADT_BA*	Cap_AB*	Cap_BA*
1: Rural Principal Arterial—Interstate	641	21.35%	129.91	9.40%	--	--	53,767.59	40,952.18
2: Rural Principal Arterial—Other	2,491	32.58%	800.64	30.51%	--	--	39,623.27	28,132.00
6: Rural Minor Arterial	2,679	27.36%	776.73	23.40%	--	--	20,633.38	17,270.75
7: Rural Major Collector	2,545	19.16%	811.8	15.67%	--	--	16,887.81	15,289.59
8: Rural Minor Collector	2,786	12.11%	1,210.24	11.41%	--	--	15,899.04	15,522.51
9: Rural Local	101	25.90%	36.64	21.43%	--	--	16,504.12	16,425.52
10: Unknown	1,711	99.36%	252.86	99.24%	--	--	11,352.33	5,691.45
11: Urban Principal Arterial—Interstate	2,109	45.51%	270.49	31.88%	--	--	68,369.33	54,330.56
12: Urban Principal Arterial—Other Freeways & Expressways	911	49.51%	119.49	38.61%	--	--	54,722.79	37,808.48
14: Urban Principal Arterial—Other	7,467	49.13%	967.86	43.40%	--	--	35,101.06	23,688.33
16: Urban Minor Arterial	8,169	44.58%	1,065.18	38.74%	--	--	23,075.86	18,192.41
17: Urban Collector	4,232	31.07%	621.74	25.85%	--	--	16,744.95	14,225.29
19: Urban Local	159	25.85%	23.41	20.01%	--	--	17,530.90	14,107.44
20: Unknown	4,379	99.59%	540.93	99.73%	--	--	12,091.63	5,604.82
91: 1-Lane Roundabout	42	72.41%	0.78	66.56%	--	--	6,338.08	1,352.68
92: 2-Lane Roundabout	13	100.00%	0.28	100.00%	--	--	10,336.01	8,121.71
98: Unknown	189	100.00%	205.26	100.00%	--	--	400,000.00	400,000.00
TOTAL	**40,624**	**34.51%**	**7,834.24**	**23.78%**				

*These are averages across all links for a functional class.

Table 7: Tennessee Interchange Summary

Type	CountsOnAll	(%)	MissingOneApproach	(%)	MissingMore	(%)	TOTAL
3-way	1,125	7.26%	4,377	28.23%	10,003	64.51%	15,505
4-way	179	0.00%	580	12.11%	4,030	84.15%	4,789
5-way	0	0.00%	0	0.00%	26	100.00%	26

Table 8: Tennessee ODME Summary

	Avg. Number of Shortest Paths
With Counts	55,055
Without Counts	52,007

2.3.2 Conservation-of-Flow Checks

The main purpose of these checks is to flag junctions that do not conserve flow (i.e., junction inflow ≠ junction outflow), based on coded count values. A secondary purpose of this check is to identify junctions with counts that are too high to be plausible for one of the approaches. Specifically, the tool is designed to perform the following checks when inbound and outbound AADT data on all approaches are available:

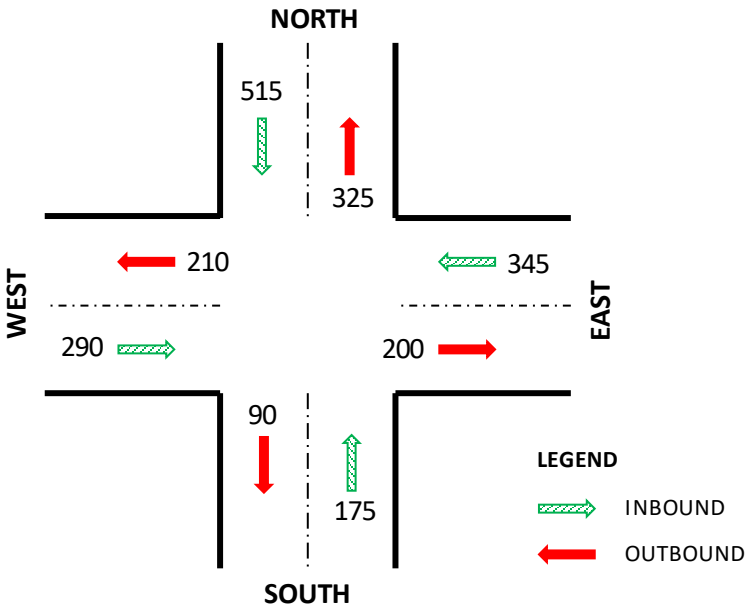
Intersection-level Check

"Total flow entering an intersection is equal to total flow exiting the intersection."

This check ensures that the conservation-of-flow holds for the intersection under consideration.

For example, at the intersection in Figure 3, the total inbound flow is 1,325 (=515+345+175+290), whereas the total outbound flow is only 825 (=325+200+90+210).

Figure 3: Intersection-level Check Example



Such an intersection is reported with a message code =1, which means that "Total flow entering the junction is not equal to the total flow exiting the junction."

Intersection Approach-Level Check #1

"Inbound AADT from a leg is less than the summation of outbound AADTs from other legs of that intersection."

Let IN_AADT_i and OUT_AADT_i be the inbound and outbound AADTs on leg i of a N-leg intersection. For each intersection approach, this check verifies whether the following relationship holds:

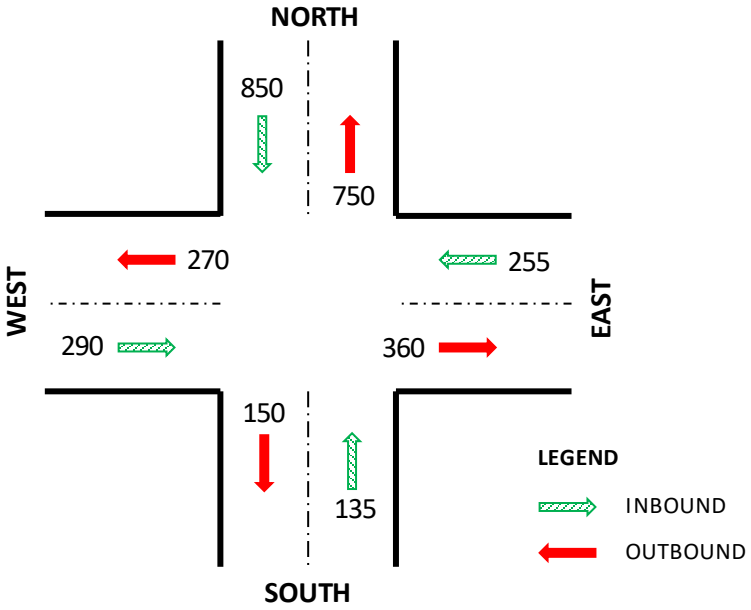
Equation 1: Intersection Approach-level Check #1

$$IN\_AADT_i < \sum_{j=1}^{N} OUT\_AADT_j, when\ j \# i$$

The aforementioned relationship is strictly an unequal relationship even though the conservation-of-flow criteria would still be satisfied with an equal relationship. This is because it is reasonable to expect that inbound flows from other approaches would contribute to the outbound flows listed. Also, it is assumed that no U-turns are allowed.

As an example, for the intersection in Figure 4, total inbound flow (1,530) is equal to total outbound flow (1,530); however, the north leg inbound flow (850) is greater than the sum of outbound flows (780) from east, south, and west legs.

Figure 4: Approach-level Check Example 1



Such an intersection is reported with a message code =2 ("Inbound flow is not less than the sum of outbound flows from other legs"). In addition, the report will include link IDs of the legs that violate the check.

Intersection Approach-level Check #2

"The ratio of inbound AADT from a particular leg and the summation of outbound AADTs from other legs is significantly less than one."

That is:

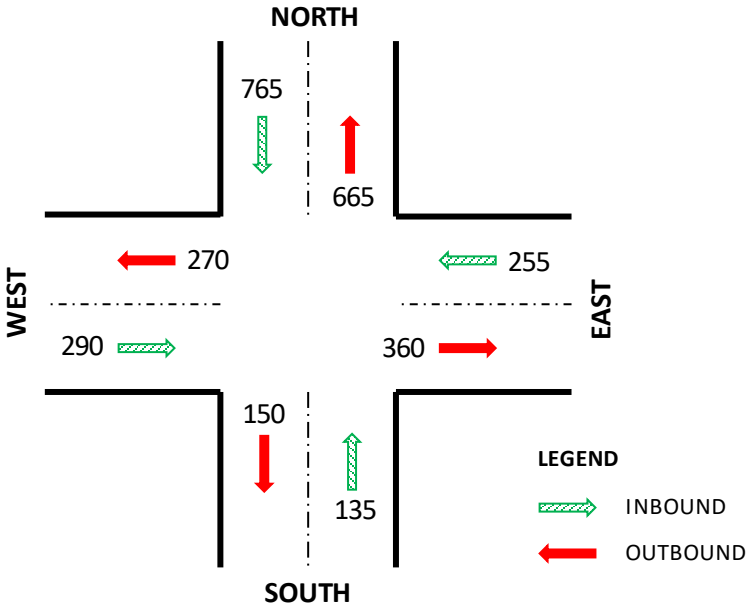
Equation 2: Intersection Approach-level Check #2

$$\frac{IN\_AADT_i}{\sum_{j=1}^{N} OUT\_AADT_j} < m, when\ j \# i$$

Where *m* <1, for the current exercise *m* was set to 0.9. The purpose of this check is to identify dominant intersection approaches, if any.

For example, in Figure 5, total inbound flow (1,445) is equal to total outbound flows (1,445). Also, inbound flow from any leg is less than the sum of outbound flows from other legs. However, the intersection does not satisfy the current check because the ratio of the inbound flow from the north leg and the sum of outbound flows from other legs (east, west, and south) is 0.981 (=765/780), which is less than the threshold of 0.9.

Figure 5: Approach-level Check Example 2



Such intersections are reported with a message code =3 ("Ratio of inbound flows and sum of outbound flows from other legs is too high"). Like the previous check, link IDs of dominant legs are also provided.

The tool generates an output ("IntersectionFlowConsCheck.csv") to report results of the above checks. The output contains only the intersections that have AADT data available on all approaches. For every intersection, a message code is attached to provide more details of the checks. The message codes reported in the outputs are described in Table 9.

Table 9: Intersection Check Output Message Codes

Msg	Label
0	Passed intersection checks
1	Total flow entering the junction is not equal to the total flow exiting the junction
2	Inbound flow is not less than the sum of outbound flows from other legs
3	Ratio of inbound flows and sum of outbound flows from other legs is too high

Intersection Turning Movement Check

In addition to the aforementioned checks, the tool also calculates turning movements and checks the compatibility of counts on approaches. For this, directional (i.e., inbound or outbound) AADTs on each approach are required. As a start, all turning movements are set to 1. For each leg, all outbound turning movements are then factored by multiplying them with the ratio of outbound flow and the sum of outbound turning movements from the leg. Next, for each leg, all inbound turning movements are factored by multiplying them with the ratio of inbound flow and the sum of inbound turning movements from the leg.

After inbound and outbound factoring, a gap value is calculated:

Equation 3: Gap Value Calculation for Intersection Turning Movements

$$Gap = \sum_{j=1}^{N}\left[1 - \left(OUT\_AADT_j \Big/ \sum_{i=1}^{N} Turn\ Movement_i\right)\right]$$

The gap value is compared against a user-provided threshold. If the value is lower than the threshold, then the current turning movements are considered final. Otherwise, outbound and inbound factoring continues until the gap falls below the threshold or until the number of iterations exceeds 200, whichever comes first. If the turn movement calculations are stopped due to iterations exceeding 200, then a message code =2 ("Turning movements cannot be calculated, please check input flows") is reported in the output file. This generally indicates a problem with the consistency of the counts.

If an intersection does not pass the intersection-level check while calculating turning movements, then a message code =1 ("Flows from one or more legs are too high to calculate turning movements") is reported.

The message codes written in the output file ("IntersectionTurnMovements.csv") are described in Table 10.

Table 10: Intersection Turning Movements Output Message Codes

Msg	Label
0	Completed
1	Flows from one or more legs are too high to calculate turning movements
2	Turning movements cannot be calculated; please check input flows

2.3.3 Junctions with Missing Data

The tool also calculates/provides guidance on missing AADT values at an intersection if all available AADT counts are reasonable but the intersection experiences the following:

- A missing inbound count.
- A missing outbound count.
- Missing approach counts (both inbound and outbound).

In addition to the preceding cases, there may be other instances where AADT data are missing for two or more intersection approaches. In such cases, it may be possible to provide rough estimates using link capacities only, though such estimates should be treated with extreme care. This version of tool does not provide any guidance for these types of intersections.

A Missing Inbound Count

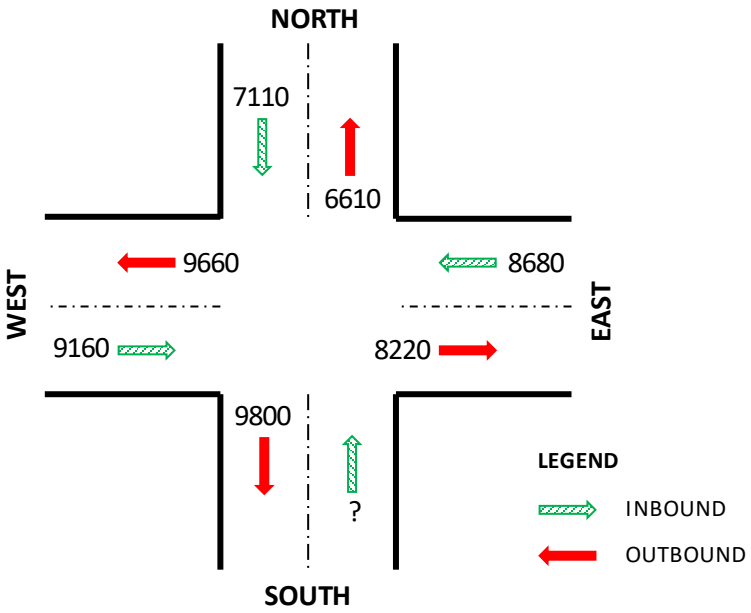
The missing AADT inbound value may be calculated when all other inbound and outbound AADT counts are available:

Equation 4. Calculate a Missing Inbound Count

$$IN\_AADT_i = \sum_{j=1}^{N} OUT\_AADT_j - \sum_{i \neq j, j=1}^{N} IN\_AADT_j$$

For example, the intersection in Figure 6 has counts in all directions except inbound from the south leg. This follows the notation adopted in the previous section.

Figure 6: Example of a Four-Leg Intersection with a Missing Inbound Count



The missing south leg inbound AADT count can be calculated as follows:

Equation 5: Calculations of Missing Inbound Count for South Leg Inbound—Example

$$IN\_AADT_S = (OUT\_AADT_N + OUT\_AADT_E + OUT\_AADT_W + OUT\_AADT_S) - (IN\_AADT_N + IN\_AADT_E + IN\_AADT_W)$$

$$IN\_AADT_S = (6{,}610 + 8{,}220 + 9{,}660 + 9{,}800) - (7{,}110 + 8{,}680 + 9{,}160)$$

$$IN\_AADT_S = 9{,}340$$

In a situation where total inbound flow from other legs is higher than the total outbound flow, then the missing AADT count will be calculated as a negative value. Such intersections are reported with a message code =2 (“Cannot calculate a missing inbound count –total outbound is less than total inbound”) in the output file.

A Missing Outbound Count

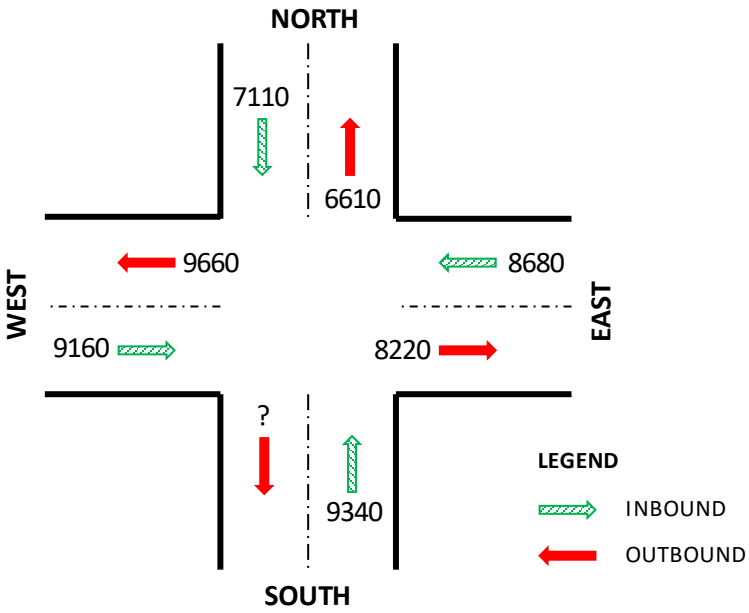
The missing AADT outbound value may be calculated when all other outbound and inbound AADTs are available:

Equation 6: Calculation of a Missing Outbound Count

$$OUT\_AADT_i = \sum_{j=1}^{N} IN\_AADT_j - \sum_{i \neq j, j=1}^{N} OUT\_AADT_j$$

To illustrate, imagine a scenario based on the previous example (Figure 6) where the south leg inbound direction AADT is available but the outbound AADT is missing (Figure 7).

Figure 7: Example of a Four-Leg Intersection with a Missing Outbound Count



The missing outbound count can be calculated as follows:

Equation 7:Calculation of Missing Outbound Count for South Leg—Example

$$OUT\_AADT_S = (IN\_AADT_N + IN\_AADT_E + IN\_AADT_W + IN\_AADT_S) - (OUT\_AADT_N + OUT\_AADT_E + OUT\_AADT_W)$$

$$OUT\_AADT_S = (7{,}110 + 8{,}680 + 9{,}160 + 9{,}340) - (6{,}610 + 8{,}220 + 9{,}660)$$

$$OUT\_AADT_S = 9{,}800$$

In situations where total outbound flow from other legs is higher than the total inbound flow, then the missing AADT count will be calculated as a negative value. Such intersections are reported with a message code =4 (“Cannot calculate a missing outbound count –total inbound is less than total outbound”) in the output file.

Missing Approach Counts

To calculate ranges of the missing inbound and outbound AADT values when count data are available for all intersection approaches except one, the missing AADT counts should observe the following relationships:

Equation 8: Range of Missing Approach Counts

$$n\times(\sum_{j=1}^{N} OUT\_AADT_j) \le IN\_AADT_i \le m\times\left(\sum_{j=1}^{N} OUT\_AADT_j\right), where\ 0 < n < m < 1\ and\ j\#i$$

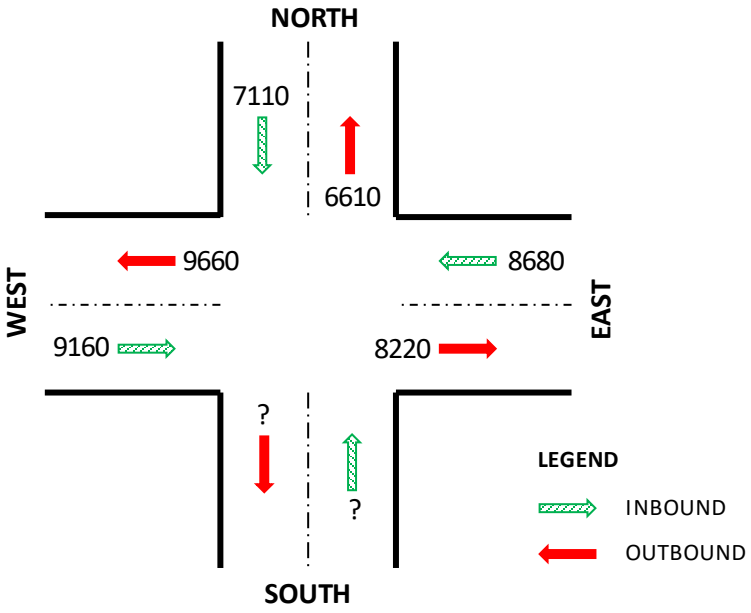
$$and$$

$$n\times(\sum_{j=1}^{N} IN\_AADT_j) \le OUT\_AADT_i \le m\times(\sum_{j=1}^{N} IN\_AADT_j), where\ 0 < n < m < 1\ and\ j\#i$$

In Equation 8, the default values of *n* and *m* were set to 0.1 and 0.9, respectively. Though, these values may be revised to obtain a tighter range if local information is available.

To demonstrate, consider the intersection shown in Figure 8.

Figure 8: Example of a Four-Leg Intersection with Missing AADT Counts on One Leg



In this example, inbound and outbound flows are available for all legs except south leg. The relations specified above indicate the following:

Equation 9: Calculate Range of Missing Approach Counts—Example

$$0.1\times(OUT\_AADT_N + OUT\_AADT_E + OUT\_AADT_W) \le IN\_AADT_S \le 0.9\times(OUT\_AADT_N + OUT\_AADT_E + OUT\_AADT_W)$$

$$and$$

$$0.1\times(IN\_AADT_N + IN\_AADT_E + IN\_AADT_W) \le OUT\_AADT_S \le 0.9\times(IN\_AADT_N + IN\_AADT_E + IN\_AADT_W)$$

That is:

Equation 10: Final Ranges of Missing Approach Counts—Example

$$2{,}449 \le IN\_AADT_S \le 22{,}041 \quad and \quad 2{,}495 \le OUT_{AADT\,S} \le 22{,}455$$

In this scenario, if local knowledge indicates that the values of *n* and *m* are closer to 0.2 and 0.4, then the revised AADT ranges are as follows:

Equation 11: Ranges of Missing Approach Counts with New Factor Values—Example

$$4{,}898 \leq IN\_AADT_S \leq 9{,}796 \ \ and \ \ 4{,}990 \leq OUT_{AADT\ S} \leq 9{,}980$$

The tool produces a separate output file ("IntersectionCalculateCount.csv") for intersections that are eligible for missing data calculations. Table 11 presents the message codes reported in the output.

Table 11: Intersections with Missing Data Output Message Codes

Msg	Label
1	Calculated a missing inbound count
2	Cannot calculate a missing inbound count: total outbound is less than total inbound
3	Calculated a missing outbound count
4	Cannot calculate a missing outbound count: total inbound is lower than total outbound
5	Calculated missing approach counts (both inbound and outbound)
6	Cannot calculate: inbound/outbound flow is not available

The intersections that have more than one approach counts missing are reported in a separate output file ("IntersectionMissingCount.csv"). The link IDs of approaches with missing counts are also provided.

3.0 Passive Origin-Destination Data

Passive OD data include information from observations of millions of individual trips that can be harnessed for travel modeling and forecasting. This section introduces and explains several types of data and how these data are processed to produce OD matrices for data-driven modeling and forecasting purposes. The section includes information that agencies can use to decide what types of data to obtain and use and how to clean and process it or understand the cleaning and processing provided by others.

3.1 Types and Sources of Demand Data

Passively collected Big Data is a rapidly evolving subject area. As recently as 10 years ago, as of 2017, there were no commercial sources of passively collected OD data. However, over the past several years, many organizations and companies have begun to offer their data for use in transportation planning and analysis. As of 2017, there are four technologies (or types) of passive OD data in use (each of these types of data are described and discussed in some detail below):

1. Cellular Tower Signaling
2. LBS (Location Based Services)
3. GPS (Global Positioning Systems)
4. Bluetooth

Agencies should consider the following when determining the best source of data.

First, determining the best data source depends significantly on how the data is to be used. Currently, the primary transportation uses of passive OD data are for developing and validating regional (including statewide) travel forecasting models and providing input demand for simulation or dynamic network models usually related to project-level design. Thus, it is important to understand if the primary interest and use of the data is at a regional level or a corridor level, as different data sources are better suited to providing data at these scales. Moreover, it is also important to understand whether what is desired is a single snapshot of travel during a period or whether the data purchase is intended to support the development of multiple datasets (e.g., to support several different design projects or corridor studies).

Second, while the primary use of OD data is to understand the pattern of trips between OD pairs, some data sources can also provide information on travel time and travel time reliability between OD pairs. While travel time is more often dealt with at the level of the roadway network, and an understanding of congestion on individual facilities is important, it is also worth considering the value of OD-level data on travel times and travel time reliability or variability. In dealing with travel time reliability OD measures may be of more interest and use than facility of 'link' level metrics for certain applications. While OD-level travel time reliability has only recently become available, it represents a potential valuable resource for future applications.

Finally, the information provided here is based solely on the experience of the authors. The authors have worked with passive data—including all four types and all major vendors/providers—in over one dozen states; however, this does not guarantee that this experience is representative or fully comprehensive. Moreover, Big Data continues to evolve rapidly and while the information presented here is believed to be accurate at the time of writing, the reader should verify whether it is still correct at the time of reading. The use of LBS data is very new at the time of this writing

and its characterizations are based on limited experience and early product releases, which later experience or releases may make obsolete.

Table 12 through Table 16 summarize and compare various aspects of each type of passive OD data currently available. The following subsections expound on key elements contained within the tables for each type of data. First, while these tables are meant to assist with understanding some of the relative strengths and limitations of these data, they should not be relied upon independent of the verification of the applicability of these generalizations to the area of interest. While there is some general basis for the characterization of data costs in the tables, the reader should not assume that a source characterized as expensive is always costlier than one characterized as intermediate. Competition has and continues to produce changes in pricing as this rapidly evolving market seeks equilibrium; different regions should always check the pricing of various data sources in their area. Similarly, sample penetration and—to some extent—locational precision are two important considerations that do vary importantly by region; it is important to verify these for an area of interest.

Second, regarding data use, Bluetooth is the only data source that does not generally support regional-level analysis or the development of travel forecasting models in general. The degree or way in which a type of data can support corridor-level analysis is described in Table 15 in the row "Select Link/Corridor Analysis" since the estimation of corridor-level demand in modeling is commonly referred to as "select link analysis." Passive OD data can support estimates of corridor-level demand in two different ways. Some types of data and data processing allow direct observation and purchase of demand patterns for a corridor. Other types of data (or processing) do not allow direct observation of corridor-level demand, but all types of data can support estimates of it using select link analysis with a network assignment model. The latter, being an indirect estimate rather than a direct observation, is clearly susceptible to the introduction of error from a network assignment model, while direct observation is not. However, in some circumstances, indirect analyses may be less expensive, and it is currently easier to address trip-length biases in indirect analyses than direct analyses.

Finally, regarding data use, most datasets provide either only total traffic or truck traffic. As a result, it is important to distinguish between applications that are focused exclusively on trucks versus and applications that are only concerned with total traffic versus applications concerned with truck and automobile traffic as separate and distinct.

Table 12: Comparison of Types of Passive OD Data (Description)*

Description	Cell-Tower Signaling	LBS	GPS	Bluetooth
Universe	All travel	All travel	Heavy trucks, medium from some providers, private from some providers	All travel
Time Periods	Average weekday or average weekend or individual day of week; multihour periods within the day	Average weekday or average weekend or individual day of week; multihour periods within the day	Generally customizable down to individual hours of the day; effort to get multiple time periods may vary significantly by vendor	Generally customizable down to individual hours of the day; effort to get multiple time periods may vary significantly by provider
OD Demand Types	Aggregate trip ODs	Aggregate trip ODs	Aggregate trip ODs; sometimes disaggregate traces also available but with restricted use	Disaggregate trip ODs
OD Travel Time Data (Including Reliability)	Not available	Not available	Available with varying degrees of processing effort depending on provider	Generally produced as part of the processing of trips

Table 13: Comparison of Types of Passive OD Data (Precision and Coverage)*

Precision and Coverage	Cell-Tower Signaling	LBS	GPS	Bluetooth
Locational Precision	>100 m often ~200–2000 m	10–100 m often ~30 m	1–10 m	10–100 m
Sample Penetration	6–10%	5–8%	9–12% truck; ~0.5% private	4–9%
Data Collection Time Period	Typically 1 month	Too new to know	1 month to 2 years depending on provider and pricing	Typically <1 month
Coverage Issues	Poor coverage in some (mostly rural) areas	--	--	Coverage limited—requires mounting detector devices

Table 14: Comparison of Types of Passive OD Data (Representativeness and Expansion)*

Representativeness and Expansion	Cell-Tower Signaling	LBS	GPS	Bluetooth
Trip-Length/Duration Bias	Confirmed	Suspected	Confirmed	Not suspected
Included/Default Expansion	Residence market share-based; generally requires adjusted to counts	None/single count-based factor, believed to require adjustment for biases	None/single count-based factor, generally requires adjustment for biases	Typically expanded to counts

Table 15: Comparison of Types of Passive OD Data (Segmentation and Applications)*

Segmentation and Applications	Cell-Tower Signaling	LBS	GPS	Bluetooth
Number of Zones	Limited by pricing and locational precision	Depends on pricing scheme	Relatively unlimited in most pricing schemes	Limited by number of detector devices
Select Link/Corridor Analysis	Generally indirect only	Indirect only currently but a subset may support direct in the future	Limited or unlimited direct depending on provider, or indirect	Direct only if detector placement allows; indirect
Filtering of Intermediate Stops on Long Trips	Premium option	Not currently available	Depending on provider may be possible as a post-process	Possible as a post-process
Residency Information	Premium options for regional residents vs. nonresidents or home block groups	Not currently available but LBS could support residence class data	Not available due to ID persistence limitations	Generally not possible
Purpose	Premium option for imputed purposes	Premium option for imputed purposes	Not available due to ID persistence limitations	Generally not possible
Vehicle Class	Not available	Not currently available	From some providers *Heavy and medium trucks, private vehicles*	Generally not possible

Table 16:Comparison of Types of Passive OD Data (Resource Requirements and Availability)*

Resource Requirements and Availability	Cell-Tower Signaling	LBS	GPS	Bluetooth
Data Cost	Intermediate	Expensive	Inexpensive to Expensive *depending on provider, amount/length of data period, and amount of processing included*	Expensive
Additional Processing Required	Intermediate	Limited to Intermediate	Substantial to Limited *depending on provider*	Usually included in price
Vendors	AirSage	StreetLight, Cuebiq	ATRI, StreetLight, INRIX, TomTom, HERE	TTI, RSG, others

**Presents generalizations based on information available at the time of writing that may or may not apply to specific regions.*

3.1.1 Cell-based Data

Cell phones regularly communicate with their networks through control channel messages. This cell-tower signaling can locate and track individual cell phones using triangulation and other inferences with signals sent between phones and towers. This was one of the first technologies harnessed to provide passive OD data on a large scale, after two of the four largest cell phone service providers in the United States partnered with a vendor, AirSage, to process and sell derived data products, including OD trip tables, based on their tower signaling information. The resulting anonymous AirSage dataset is based on data from over 100 million devices and provides coverage for most areas in the country (although there are gaps in some, particularly rural, areas). Disaggregate, cell-based data is not available. Data are drawn from cell phone users, and this is generally assumed to represent the adult traveling public—including truckers, who cannot be separately identified.

Precision and Coverage

Cell-based OD matrices can be obtained for most time periods of potential interest, including average weekday, average weekend day, individual day of the week, and multihour periods within the day. The spatial resolution or precision of the data is limited. Locations are generally only known with a precision of more than one hundred meters and sometimes only within one to two kilometers in areas of limited tower coverage, but precision tends to be better in urban areas with better tower coverage. Cell-based data is typically purchased in observation / data collection periods of one month, although sometimes multiple months of data are purchased and often some discount is available for purchases of multiple months.

Sample penetration can vary significantly depending on the market share of various cell phone service providers and, as noted previously, there are some areas with no coverage at all. The authors have found these samples typically include approximately 6–10% of vehicles in a corridor. These samples may include observations from a significantly larger portion of the population, perhaps as much as 30% or more depending on service provider market shares. However, not

all trips by a person are necessarily observed. Therefore, the portion of trips observed is less than the portion of the population included in the sample. These figures vary by region, and some regions may have even larger samples than this range while others (especially rural areas) may not achieve this level of penetration.

Representativeness and Expansion

As with most types of data passively collected from mobile devices, the frequency of positional observations varies within the dataset. In the case of cell-based data, the frequency of signaling between the cell phone and the tower can vary significantly based on the tower technology, the individual make and model of phone, the phone's operating system and settings, and the usage of the phone. In some circumstances, a phone could be communicating with towers every few seconds; in other cases, a phone may go one hour or more without communicating with a tower, particularly when the phone is not in use. Infrequent observations of position lead to the omission of some trips in the data; the odds that a trip is omitted decrease as the duration of the trip increases. The result is a systematic bias in the data in which longer trips are over-represented relative to shorter trips. (14) This trip-length bias has also been observed in GPS data (7,8) and is suspected in LBS data. Although the significance of the bias and the precise details of how it arises vary somewhat in the different types of data, it appears to be a general problem in passive data that arises from varying or infrequent observations. Failure to account for such biases can lead to erroneous representations and faulty predictions of trip lengths, trip flows between origins and destinations, and present and future travel activity patterns in general. The varying frequency of observations also prevents the development of OD travel time or reliability metrics from cell-based data.

Cell-based data is typically pre-expanded based on proprietary estimates of service provider market share at imputed residence locations. This residence- or population-based expansion can be helpful in addressing biases related to market shares. However, this type of expansion does not address systematic biases that can arise in the data when people travel to and from locations with poor coverage or when the trip-length bias arises from the varying frequency of observations. Therefore, it is important to correct the expansion of cell-based data to address these systematic biases. This is generally done by developing expansion adjustment factors based on traffic count data as discussed in Section 4.0 of this volume.

Segmentation and Applications

Cell-based data is frequently used to support regional applications such as modeling. Given its limited spatial precision it can only provide direct observations of facility or corridor-level demand under special circumstances, such as rural interstates with no nearby parallel corridors. However, cell-based data can be used in conjunction with a network assignment model and select link analysis to provide estimates of corridor-level demand.

The size and the number of zones within an area is also limited by the spatial resolution of the data, and pricing considerations also make more zones costlier. For these reasons, cell-based data often cannot be obtained for all zones in a regional model; aggregation or grouping of some zones into districts is commonly required. As a rough rule of thumb the spatial resolution of cell-based data approximates Census Block Groups. Like block groups, the resolution is better and can support smaller zones in denser urban areas versus more rural areas. The number of block

groups in a region is a starting point when estimating the maximum number of zones cell-based data might support in a region. However, this is just a rough rule of thumb and starting point for understanding the precision of cell-based data. These data can vary substantially between and within regions because it depends on cell tower locations; it is important to verify and understand the precision cell-based data can provide for a region of interest.

Cell-based data can support long-distance and visitor travel analysis and modeling and is better suited to this task than GPS data. Unlike GPS data, device IDs are persistent for a month or more in cell-based data. Long-distance, multiday, and short-distance "visitor" trips made outside of the traveler's home region can be identified reliably. Moreover, for many types of long-distance travel analysis, it is important to understand travelers' "true" destinations rather than intermediate stops (e.g., for food, fuel, rest) that they make along the way there. Cell-based data can be processed by the provider, for an additional fee, to provide this filtering of intermediate stops to better understand long-distance travel patterns. Correcting for trip-length biases in the data is especially important in datasets that include both long- and short-distance trips.

Information on travelers' residence or home location, which is supported by ID persistence in cell-based data, can also support the imputation of trip purpose (e.g., whether a trip is to or from the traveler's home or work location). However, several studies have shown significant differences between imputed purposes and reported purposes, leaving the accuracy of imputation methods in question. (9,10) One important source of difficulty in the imputation of purposes and difference between imputed and reported work locations is that imputation generally assumes that the place a person spends most their day at is that person's workplace. This classifies many students and volunteers as workers and their schools or volunteering locations as workplaces. Homes can also be misidentified as workplaces for third-shift workers. As an alternative to purpose imputation, Census data on commute flows can be used to segment cell-based trips into work and nonwork. Vehicle class and travel model currently cannot be imputed or observed in cell-based data. However, cell-based data can be broken into truck and nontruck segments in combination with truck GPS data. (For more information, see Section 3.4.)

3.1.2 LBS Data

LBS data is aggregated from smartphone and other mobile device applications ("apps"). The LBS data is not based on a single technology such as cell-tower signaling or GPS; rather, these data represent the best location available to mobile apps, which could come from GPS, Wi-Fi, Bluetooth beacons, or cell-tower signaling under various circumstances (although its reliance on the last is limited). Most LBS locational data comes from Wi-Fi beacons and GPS. LBS data is the newest type of passive OD data and has only recently become available for widespread use in transportation analysis as of 2017. Currently, the main provider of LBS data in the United States claimed its sample was drawn from over 160 major mobile apps with over 50 million users in the United States. Like cell-based data, privacy considerations limit data to aggregate trip OD matrices. Data are drawn from mobile internet device (i.e., smartphone and tablet) users, and this is generally assumed to represent the adult traveling public—including truckers, who cannot be separately identified.

Precision and Coverage

Also, like cell-based data, LBS OD matrices can be obtained for most time periods of potential interest, including average weekday, average weekend day, individual day of the week, and multihour periods within the day. More fine-grained temporal resolution down to the hour may be available, but its reliability is not known.

In contrast to cell-based data, LBS data offers better spatial precision, although its resolution is less than what is available with GPS data. Locational precision is generally between 10 and 100 meters, with most data observations precise to better than 50 meters. Precision exceeds that of cell-based data due to the availability of multiple technologies to provide locational information.

Sample penetration can vary by region due in part to the varying popularity of apps in different markets. However, sample penetration is expected to be less variable than in cell-based data given the large number of apps LBS draws on. Based on a limited number of observations, the authors have found LBS data to include 5–8% of the vehicles in a corridor. The sample is believed to include up to 15% of the population, but with varying frequency of observation both between individuals and for individuals over time depending on app usage. This current sample penetration appears—based on a limited number of observations—to be like (but slightly lower than) cell-based data (this generalization will not apply to all regions) and substantially higher than GPS data (for total or non-truck travel). Moreover, the LBS data sample has steadily increased, and it is possible its sample penetration may be higher soon. Therefore, as emphasized previously, it is important to verify sample penetrations in each region at the time of data acquisition.

Representativeness and Expansion

The frequency of locational observations varies within the LBS dataset. This matches most other types of passive OD data and is expected to lead to a systematic bias related to trip length or duration (this has not yet been confirmed in LBS data since it has only recently become widely available). Trip-length biases in LBS data may be less than in cell-based data, but these biases are still expected to be significant. As noted previously, failure to account for this type of systematic bias can lead to erroneous representations and faulty predictions of trip lengths, trip flows between origins and destinations, and present and future travel activity patterns in general.

LBS data is currently not typically pre-expanded, so users must expand the data. The data provider may provide a tool for scaling the data based on the average ratio of data observations to traffic counts. However, such simple scaling does not address systematic biases such as those related to trip length or frequency of observation. Therefore, data expansion involving a system of multiple expansion factors based on traffic counts is recommended (see Section 4.0 for more information).

Segmentation and Applications

LBS data is well-suited to regional applications such as modeling. The locational precision of LBS data makes it better able (in theory) to support facility- or corridor-level applications than cell-based data; however, it is not as reliable for this purpose as pure GPS data. Currently, like cell-based data, direct observations of corridor-level demand from LBS data are not available. However, this may change in the future and a subset of more precise LBS observations may function like GPS data to provide direct estimates of corridor-level demand. Regardless, like cell-based data and all passive OD data, LBS can support indirect estimates of corridor-level demand using a network assignment model and select link analysis.

The size of (and the number of zones within) the region that LBS data can support is large and LBS data can generally provide data for a regional travel model's zone system. Some pricing schemes may involve the number of zones as a factor in the cost, but the predominant pricing scheme to date has offered an unlimited number of zones, with the price varying based on the population of the region.

Like cell-based data and unlike most GPS data, LBS data have longer term device ID persistence. In theory, this could be used to support long-distance and visitor travel analyses. However, as of 2017, this was not yet a standard option offered by data providers. Imputation of purpose (like with cell-based data) is not currently offered, but is expected. Although not yet verified, it is expected that the accuracy of imputed purposes would be like that of cell-based data due to similar imputation algorithms. It is recommended that imputed purpose information be used with caution until the accuracy of imputed purposes can be verified against reported purposes.

Vehicle class and travel model currently cannot be imputed or observed in LBS data. However, like cell-based data, LBS data can be broken into truck and non-truck segments through combination with truck GPS data. (See Section 3.4 for more information.)

3.1.3 GPS Data

GPS data is derived from on-board vehicle devices or integrated systems, personal navigational devices, and (in some cases) personal mobile devices. GPS was one of the first technologies used to provide passive OD data on a large scale, specifically for truck travel. Some GPS datasets are still specific to trucks—even heavy or multiunit trucks—while other datasets provide some data on medium-duty trucks and noncommercial or private travel. Like other types of passive OD data, GPS data is often purchased or processed to produce aggregate trip OD matrices. However, in some cases, providers may share disaggregate GPS trace data, but only with significant limitations on its use. One common consideration in whether disaggregate GPS data can be obtained is whether the organization obtaining the data can enter into a binding nondisclosure/data-sharing agreement. Government agencies and universities that are subject to "sunshine" laws are sometimes precluded from access to this level of the data, whereas private consulting companies may not be. However, while consulting firms may obtain access to the disaggregate data, they are generally prohibited from sharing these data with public agencies; instead, these companies can often only provide aggregate data products, model parameters, and similar.

Because of its high level of precision in both space and time, GPS data can provide not only trip OD matrices but also OD travel time metrics, including OD travel time reliability. Although less commonly used, this information can be valuable for modeling and many travel analyses. For instance, in the Tennessee statewide model case study, OD travel times from truck GPS data were used to validate the model's skims.

Unlike some other types of data, there are multiple providers of GPS data. The characteristics of the data products offered by providers vary. The principal difference is often the amount of processing done by the provider versus the amount of processing that is left to the user. Some providers offer data at a lower price, leaving most processing to the user. This can offer flexibility in how the data is processed, and for some applications this can result in a better final data product

or lower total costs even after allowing for the cost to process the data. However, in other cases, a data product or platform with built-in processing may cost more than raw data but less than the combined cost of raw data and required processing. It is therefore important to understand the full cost of data and processing required for an application in determining the most cost-effective source of GPS data.

Precision and Coverage

GPS OD matrices are available for most time periods of interest, including average weekday, average weekend day, individual day of the week, and even down to individual hours of the day (or possibly less). Different providers offer default observation/data collection periods that can vary from a single month to multiple years.

GPS data is the most precise source of locational data. Precision is generally in the range of 1 to 10 meters, and often less than 5 meters. This level of precision allows vehicles to be located not only in zones at their origins and destinations but to individual roadways along their routes.

The sample penetration varies significantly by vehicle class. Multiple providers can offer sample sizes generally in the range of 9–12% of trucks on the road. Private, non-truck samples are currently quite small, however (0.5% or less of the total traffic in a corridor). This current low sample size for personal travel significantly limits the degree to which it can be generalized and expanded to accurately represent all travel. However, the non-truck GPS sample penetration may increase and allow it to become a more valuable dataset as vehicle technology advances and the fleet becomes more connected.

Representativeness and Expansion

Truck GPS data was the first type of data in which the systematic trip-length biases were observed. (7,8) While infrequent observations play some role as in other types of passive data, this is believed to be somewhat less a factor in truck GPS data. The composition of the vehicle sample may also contribute to this bias as anecdotal information suggests that the sample is skewed toward long-haul trucks. As noted previously, failure to account for this type of systematic bias can lead to erroneous representations and faulty predictions of trip lengths, trip flows between origins and destinations, and present and future travel activity patterns in general.

Like LBS data, GPS data is currently not typically pre-expanded, so the user must expand the data themselves. The data provider may provide a tool for simply scaling the data based on the average ratio of data observations to traffic counts. However, such simple scaling does not address systematic biases such as those related to trip length or frequency of observation. Therefore, data expansion involving multiple expansion factors based on traffic counts is recommended. (See Section 4.0 for more information.)

Segmentation and Applications

Truck GPS data is well-suited to regional applications such as modeling. Truck GPS data's high-fidelity locational precision also supports facility- or corridor-level applications. Non-truck GPS data likewise has the locational precision to support both types of applications, but sample penetration currently limits its usefulness. The effort required to process GPS data for corridor-level analysis can vary significantly by provider. A higher-cost provider may also provide tools to simplify the processing of the data for these purposes, whereas a lower-cost provider may provide

data that can support this type of analysis but that requires substantial processing, particularly to obtain results for multiple corridors.

Device or vehicle ID persistence varies among GPS datasets. Some truck GPS datasets have significant ID persistence while other truck GPS datasets (and all currently available non-truck GPS datasets) have device persistence of 24 hours or less. Limited device ID persistence significantly limits the usefulness of data in understanding long-distance (or visitor) travel patterns. The filtering of intermediate stops on long-distance trips is also of importance in many applications. For trucking, the filtering of intermediate stops is important for GPS data to be compared to or combined with commodity flow data. Some truck GPS providers provide data that allows for this type of processing, whereas others currently do not. These dual, related issues pertaining to long-distance travel are of particular importance for statewide or intercity applications. They can still have some effect, but are generally much less of an issue, at the metropolitan scale.

The limited ID persistence in GPS datasets also prevents the imputation of trip purposes. However, GPS data, unlike other data types, typically provides information on vehicle class, with data exclusively on heavy trucks or broken out between heavy- and medium-duty trucks and private/personal vehicles.

3.1.4 Bluetooth Data

Bluetooth devices and technology provide another source of passive OD data. However, in many ways, Bluetooth data differ from other types of passive OD data. It is arguable whether the label of Big Data is appropriate for Bluetooth datasets as these are typically much smaller than other datasets. This is because they typically include data for only a limited number of detector locations and only for a relatively limited period of observation (days or weeks vs. months or years). Moreover, while Bluetooth technology is passive insofar as it does not require interaction with or the active cooperation of the travelers observed, it is still active insofar as it requires the placement and monitoring of detector devices. As a result, there are no nationwide vendors of preexisting Bluetooth datasets, but rather various firms capable of collecting Bluetooth data.

Like other types of data, Bluetooth data can be processed to provide trip OD matrices. But unlike other data sources, the disaggregate data are also generally available. However, Bluetooth origins and destinations are detector locations. These detector locations are not typically true trip origins or destinations in the sense of activity locations, but rather origins or destinations for specific types of travel analysis, such as where trips enter or exit a corridor or a cordon area. Bluetooth data are typically anonymous and frequently include limited data on routing, but these data can be designed to provide important high-level routing information in some circumstances. Moreover, travel times are also often processed as part of the processing of trip information, so OD travel times and sometimes even OD travel time reliability are produced as byproducts of the production of trip OD matrices.

Precision and Coverage

Like other data types, Bluetooth OD matrices can be obtained for most time periods of interest, including average weekday, average weekend day, individual day of the week, and even down to

individual hours of the day (or possibly less). Data collection/observation periods vary, but these can range from several days to several weeks—longer periods of data are sometimes collected.

Bluetooth data are generally of intermediate locational precision—similar to LBS data—in the range of 10 to 100 meters. However, careful placement of directional detectors can often fairly accurately determine location (at close to 10 meters) to identify vehicles as being in the lanes in a particular direction on a facility. In practice, this means that Bluetooth's precision can sometimes approximate GPS's precision.

The sample size of Bluetooth data varies by region, but it is generally between 4–9% of vehicles on the road. This sample size may increase over time as newer, Bluetooth-enabled vehicles comprise a larger portion of the fleet.

Representativeness and Expansion

The systematic biases related to trip length are not suspected in Bluetooth data. This differs from other types of data and is attributable to the fact that observations—rather than being opportunistic or event-based—are controlled using detector devices. There is some concern that Bluetooth samples may be skewed toward higher-income segments of the population with new vehicles with more technology options; however, there is little research to support or refute this claim.

Also unlike other data types, Bluetooth data is typically expanded to travel counts as part of the data processing. Multiple expansion factors are typically used during this process and commonly used methods are believed to produce representative data.

Segmentation and Applications

Given the unique characteristics of Bluetooth data, it is typically used for distinct applications. Unlike other data types, it is not well-suited to regional applications and modeling as it can only support a limited number of zones as a detector device is required for each zone. However, Bluetooth data is particularly well-suited to corridor studies and understanding through (external-external) trip demand in a region.

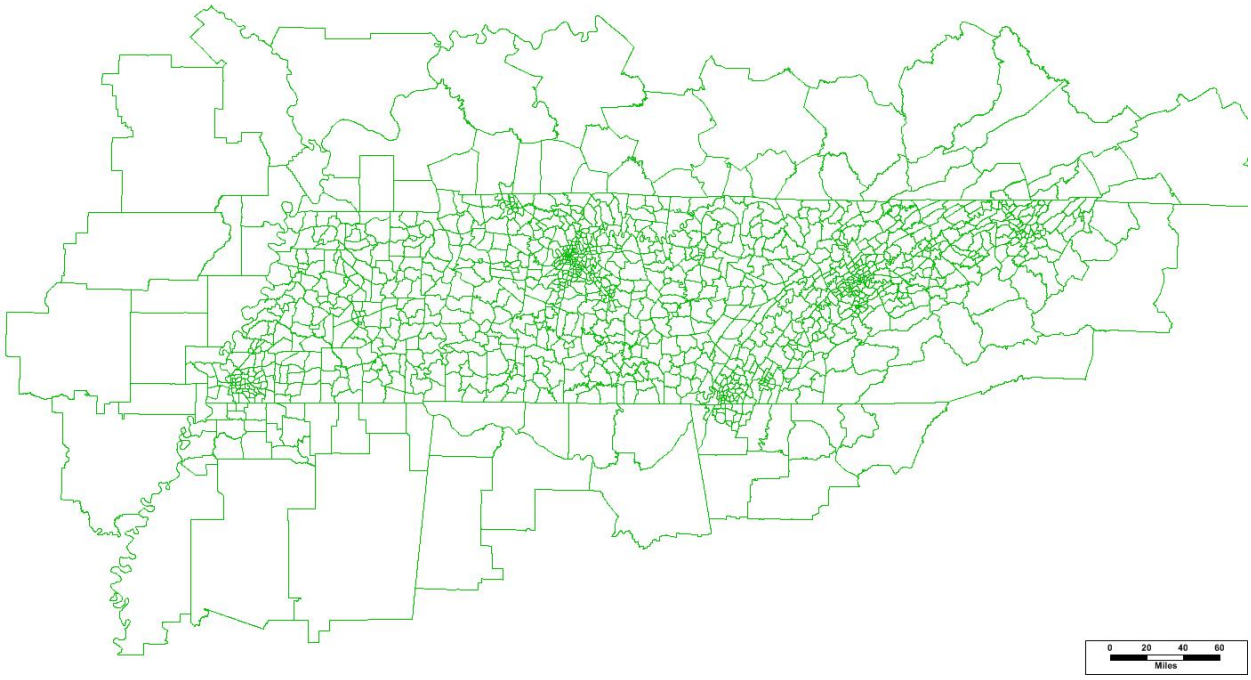
Unlike other data types—in which intermediate stops must be filtered out on long trips—Bluetooth data often excludes intermediate stops by design, although this varies. In general, Bluetooth can support both long-distance corridor studies and short-distance corridor studies.

Understanding of travelers' residences, or which vehicles represent visitors to a region, is not possible with Bluetooth data, nor is the identification of vehicle classes or the imputation of travel purposes.

3.2 Data for the Tennessee Statewide Model

Given the statewide scale and nature of the application and the importance of long-distance travel and commodity flows in this context, both cell-based and truck GPS data with ID persistence were selected and obtained to support the development of the TSTM.

Figure 9: AirSage Districts in Tennessee and Surrounding Areas



TDOT acquired O-D data from AirSage for the State of Tennessee and a halo area surrounding it. The districts used are displayed in Figure 9 and include 1,223 districts with 1,092 of these districts in Tennessee. These districts maintain approximately a 3:1 ratio with the statewide model's zones. The average population of districts within Tennessee was approximately 5,800—making them (on average) slightly larger than Census tracts. The districts had an average size of just over 35 square miles with some rural districts over 100 square miles and a handful of the smallest urban districts around one square mile. This level of resolution was selected both to ensure it would be well supported by the limited spatial precision of the data, especially in rural areas, and to achieve a feasible cost. AirSage provided coverage for all but one small rural district in Tennessee, although there was one area outside Tennessee that lacked coverage.

TDOT also acquired truck GPS-based OD data from ATRI through a consulting firm that also processed the data. An eight-week sample of trucks passing through Tennessee was obtained, drawing two weeks of data from each of the four quarters of 2013. The two-week periods were chosen to avoid major holidays; this captured some seasonal variation without holiday-specific patterns. The ATRI data included information from over 234,000 individual trucks on over 6.5 million truck trips representing approximately 11% of the trucks on the road for 56 days observed.

3.3 Data Processing

The amount and type of processing required of passive OD datasets varies depending both on the type of data and the amount of processing done by the data provider. Although they can be easily overlooked, data processing can have important effects on the data and how it can be used. All passively collected data require some form of expansion to represent all travel or all travel within a category (e.g., all truck trips). Section 4.0 of this volume is dedicated to data expansion,

particularly using traffic counts. The remainder of this section is devoted to the other types of processing that can be required for these datasets.

Many data providers include some (or even all) the processing as part of their data products. Typical processing steps are presented below so that users will be familiar with and understand the type of processing that is required, even if they do not have to perform it themselves. The processing of the truck GPS data from ATRI for the Tennessee statewide model is used as an example since it includes the least preprocessing by the provider. Some processing details presented here may differ from the processing other datasets or provided by other vendors. The processing of the ATRI data for Tennessee is provided as an illustrative example only. If a vendor omits one of the checks described here and instead performs other checks this may still be perfectly appropriate in various circumstances. Differences in the locational precision, ID persistence, and other details of the various types of data can and do make different types of data validation checks possible, impossible, or more or less helpful.

3.3.1 Identifying Stops and Trips

One of the challenges associated with passively collected OD data is identifying trip origins and destinations. In this context, the definition of a trip or a trip origin or destination is not to be taken for granted. In some contexts, particularly for detailed demand simulation modeling, origins and destinations may be desired as a latitude-longitude location or land parcel or "microzone" (like a Census block) at which a non-travel activity occurs. However, only some types of data support this level of precision and most applications do not require it. In most cases, an origin or destination is defined as a zone or district for travel analysis in which non-travel activity occurs. Trips occur between these origins and destinations to allow the traveler to participate in the activities at those locations. As fundamental and seemingly elementary as this concept or definition is, data typically does not include observations of non-travel activities, but only observations of location or position at various points in time, so the locations of activities or trip origins and destinations must be inferred. Moreover, in some cases, such as long-distance travel analyses and modeling, alternative definitions of trip origins or destinations may be preferable. (This is discussed further in Section 3.4.)

The inference or identification of trip origins and destinations or stops can be made using various criteria. Different approaches may be more amenable to data with different levels of locational precision and frequency of observations. Greater locational precision and frequency of observation allow more clear-cut criteria while imprecise or infrequent data may require more complex inferences using clustering, among other methods.

In the case of the Tennessee ATRI dataset, the GPS technology provided relatively precise locations and relatively high frequency of observations. However, for purposes of protecting privacy, the precise latitude-longitude locations from GPS were not provided. Rather the latitude and longitude were replaced by the zone in the GPS trace records, but not before the precise GPS locations were used to compute the geodetic distance between each pair of "ping" or location records. In addition to anonymous device IDs, the zone, and the distance from the last location, the original "ping" or location records also included a date-time stamp.

A data management and analysis software package was then used to process the data to produce a new data table in which each record represented a movement with an initial or "from" zone and a subsequent "to" zone and the distance, time, and the resulting computed speed. The movement data table was then processed to identify "true" trip origins and destinations and to aggregate multiple movement records into trip records. A simple algorithm was used to label each movement record as either "moving" or "stopped." The algorithm labeled records as stopped once a vehicle's speed was 5 mph or less for 5 minutes or more. (It is important to not require the speed to reach zero because even with the relatively high locational precision of GPS data, subsequent observations can appear to show small differences in location when the device has in fact remained stationary, which results in erroneous speeds.)

Additional logic was used, for instance, to prevent very small movements, such as the repositioning of a truck within a single site between a holding area and loading docks (as shown in Figure 10) from being recorded as trips. This simple algorithm will result in missed very short stops of less than five minutes, but multiunit trucks (such as those included in the ATRI dataset) rarely make stops of this duration. More recent algorithms for identifying stops often also include the change of direction or heading as part of the criteria, and there is some evidence that these more complex criteria may yield superior results. However, the simpler criteria worked well for the Tennessee dataset.

Figure 10: Brief Repositioning Movements to be Distinguished from Trips

Table 17 shows an example of a small number of records of truck GPS trace data, processed into movement records as described above. Table 18 shows the resulting trip records from the processing of Table 17. From this format, it is easy to aggregate trip records into a traditional OD matrix format.

Table 17: A Small Sample of Movement Records from Truck GPS Data

From Zone	To Zone	Distance	Time	Elapsed Time	Speed	Status
10	101032	66.0	57.7	57.7	68.8	Moving

From Zone	To Zone	Distance	Time	Elapsed Time	Speed	Status
101032	101033	16.3	14.3	72.0	68.8	Moving
101033	101015	26.6	27.9	99.9	57.5	Moving
101015	101015	0.0	5.0	5.0	0.0	Stopped
101015	101015	0.2	2.7	7.7	4.9	Stopped
101015	101015	0.3	9.8	17.5	2.0	Stopped
101015	101015	0.1	0.3	0.3	28.2	Moving?
101015	2035	37.1	60.0	60.3	37.1	Moving
2035	18099	67.8	65.4	125.7	62.2	Moving
18099	27006	5.9	5.4	131.1	65.3	Moving
27006	18023	10.0	15.9	147.0	37.8	Moving
18023	18023	0.0	5.0	5.0	0.0	Stopped

Table 18: Resulting Trip Records from Processing Movement Records

Trip	Origin	Destination
1	10	101015
2	101015	18023

3.3.2 Data Validation and Cleaning

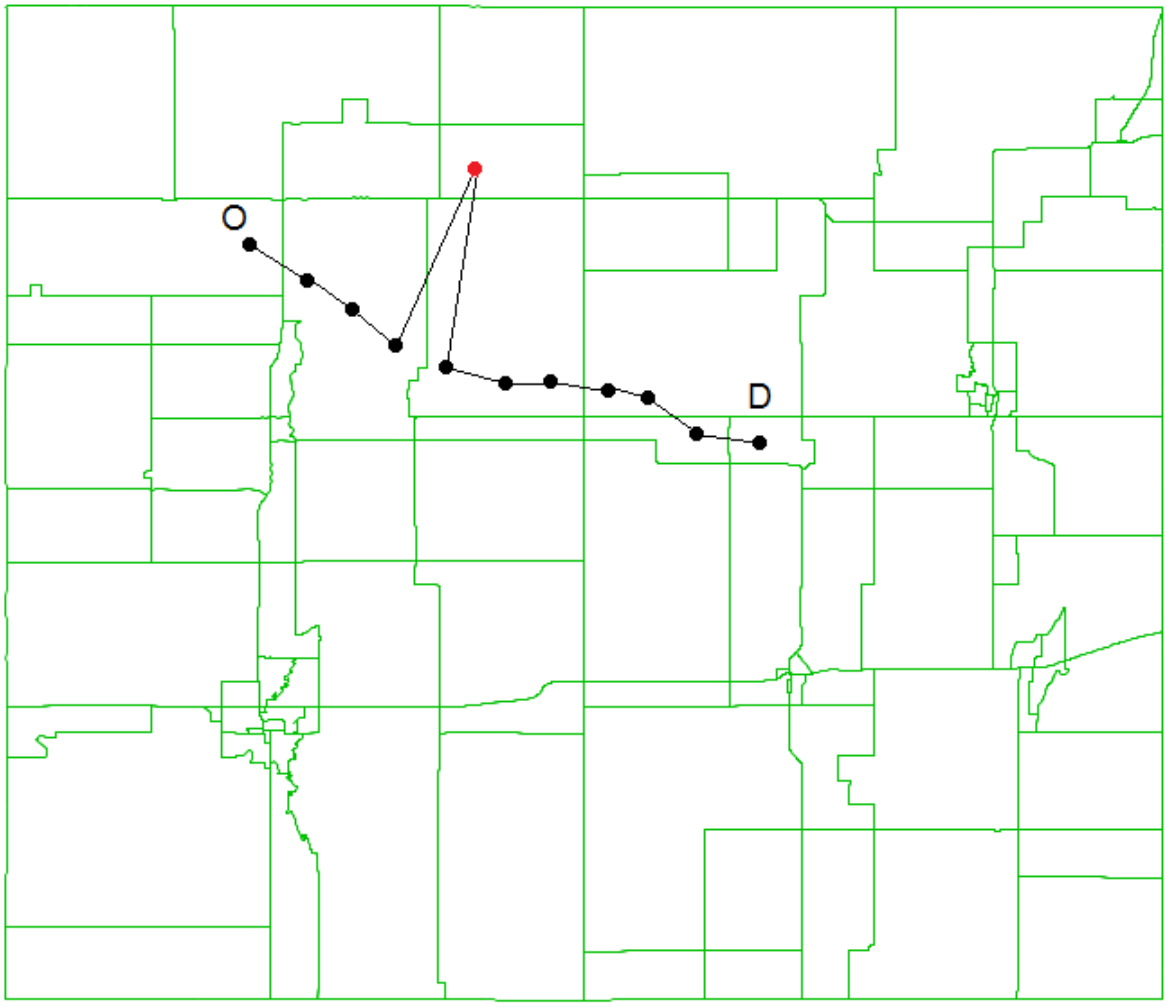
The resultant trip table must still be filtered or cleaned to ensure a valid final dataset. Various filters must be applied to the raw trip table to remove outliers or spurious data; these outliers are often due to GPS positional errors or "blips." The following sections detail the raw trip table cleaning process for the Tennessee truck GPS data, but the reader should keep in mind that the most appropriate cleaning and data validation checks will vary with the dataset.

Position Errors

Raw trip tables often include position errors or "blips." Blips refer to situations in which the GPS location jumps from one location to another one in a short time representing unrealistic movements. For example, if a record shows a movement of 40 miles in less than one minute,

then it can be safely identified as a blip and removed from the data. Although it might be ideal to attempt to recover a portion of these records, this is often unnecessary due to the large sample size. Even if errors result in discarding a few percent of the data, the clean dataset will still contain millions of records.

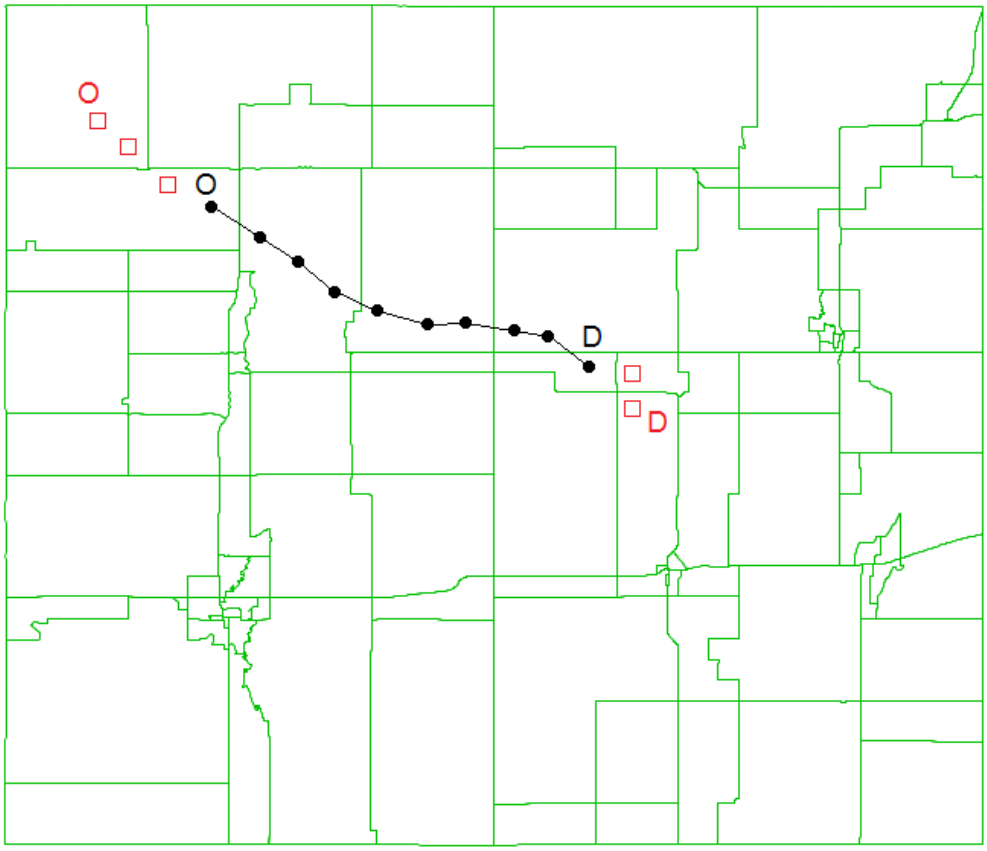
Figure 11: Example of GPS Positional Error or "Blip"



Study Period Boundaries

Some trips at the beginning or the end of each two-week data collection/observation period might be only partially represented. To avoid recording these partial trips as if they were complete, these trips are removed from the trip table. For example, if a truck is first observed in motion within a window of the global start time, then the records associated with it are filtered out since there is no record indicating its first stop or true origin. Similarly, if a truck does not show any final stop prior to the global end time, then its final trip is removed.

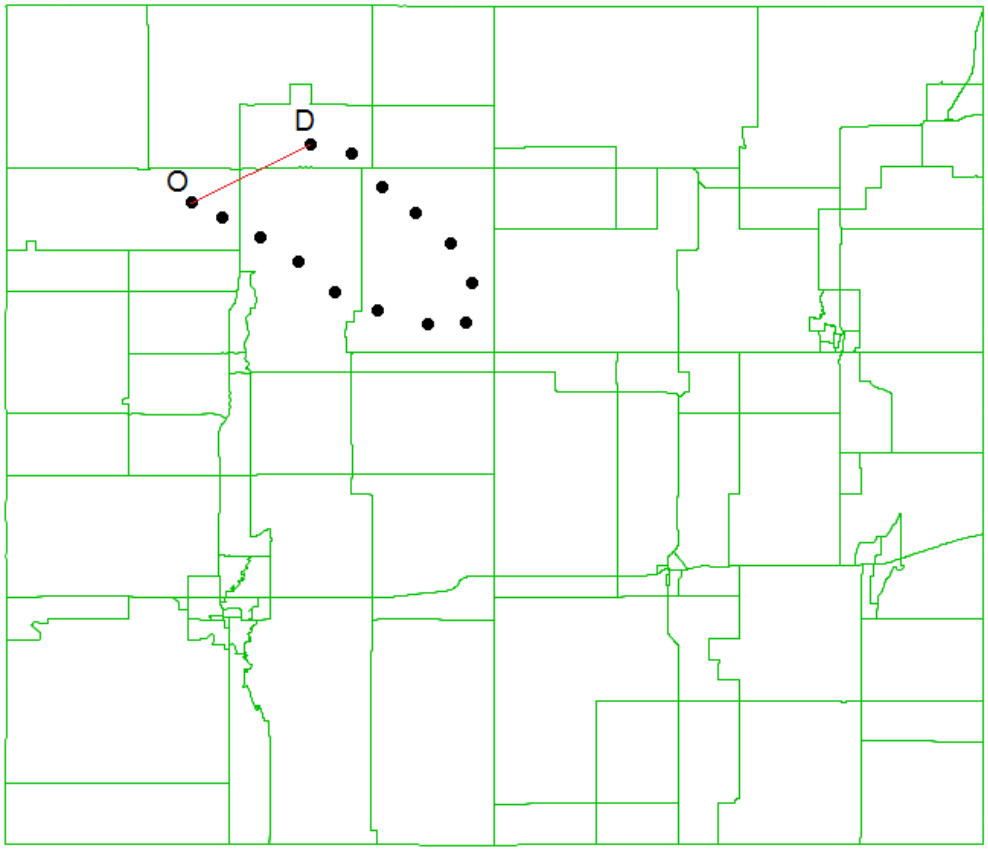
Figure 12: Truncated Trips at Observation Start or End Times



Internal Circuity

Another filter is applied if the ratio of GPS-calculated length to centroid-to-centroid geodetic distance for a trip is higher than 2.25. This internal circuity filter captures both blips that pass through the first filter and undetected stops. In the Tennessee dataset, fewer than 1% of all trips were removed using this filter.

Figure 13: Internal Circuity



3.4 Combining OD Datasets

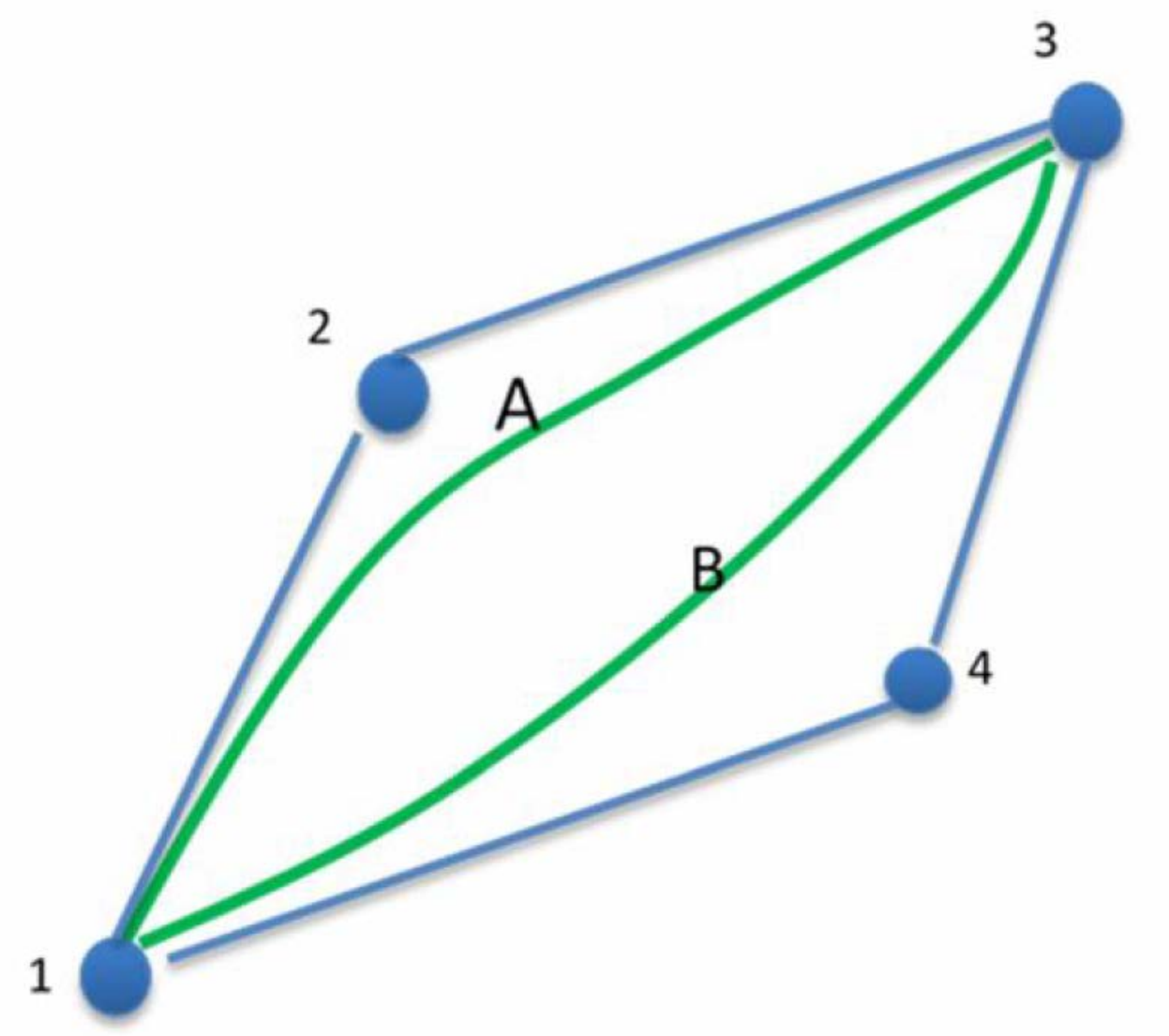
It may be necessary to combine multiple sources of OD data in cases that require distinct information on both automobile and truck traffic. Some datasets provide only total traffic and other datasets provide only truck traffic; no currently available source provides a large sample of both automobiles and trucks with the ability to distinguish between the two.

The development of the Tennessee statewide model required combining datasets. One of the key purposes of obtaining passive OD data was to support the calibration of the National Long-Distance Passenger Travel Demand Model for Tennessee. However, no dataset available at the time could provide a large sample of long-distance passenger trips excluding truck trips. Therefore, cell-based data on total travel from AirSage was combined with truck GPS data from ATRI to develop separate passenger and truck OD matrices. It is expected that similar exercises to fuse or combine multiple datasets will become more common as the availability and use of passive OD data grows.

The initial plan for the Tennessee statewide model was to simply subtract the truck trips from the truck GPS data from the total cell phone based trips. However, the initial attempt to do so revealed that there were more truck trips than total trips for 11% of the OD pairs observed in the cell phone data. Although only 0.2% of the total cell phone trips were involved, given the large number of OD pairs, this was considered a problem.

It eventually became clear that the primary reason for this was a difference in the way the two datasets were processed relative to the definition of stops or trips and long-distance trips. The Tennessee AirSage data had been purchased with filtering to remove intermediate stops (such as for fuel, meals, rest, etc.) on long-distance trips. Based on AirSage's description of their methodology, if a traveler went 50 miles from home, then the criteria for defining a stop changed. Rather than being based on the amount of time the traveler spent in the same place, a stop was instead coded only when the traveler reached the point furthest from home and began traveling back toward home. As a result, intermediate stops (stops #2 and #4 in Figure 14) between home and the assumed destination at the farthest point from home (stop #3 in Figure 14) were removed from the dataset and long-distance trips were only reported to and from home (trips A and B in Figure 14).

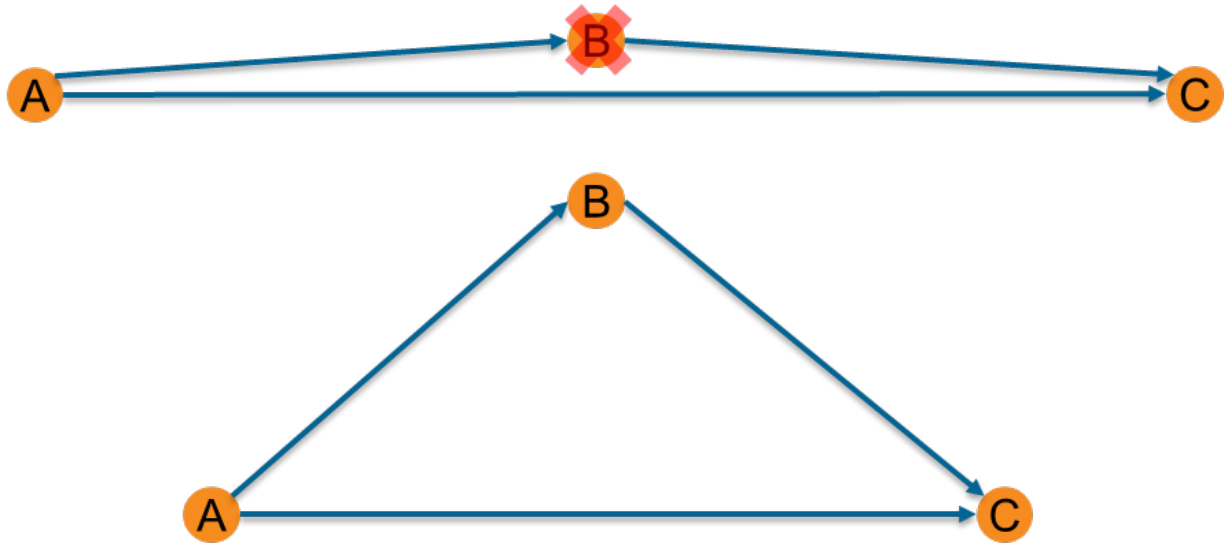
Figure 14: AirSage's Long-Distance Trip Filtering Method (source: AirSage)



In contrast to the AirSage data, the ATRI trip table originally included intermediate stops on long-distance trips. It was therefore necessary to reprocess the ATRI data, filtering out intermediate

stops. However, it was not possible to use the same filtering algorithm because “home” locations could not be reliably identified for many trucks. Moreover, it was deemed important and desirable to allow for multiple destinations on a long-distance tour (e.g., a truck carrying one shipment from Nashville to Knoxville may then pick up another shipment and take this to Chattanooga before returning to Nashville) even though these complex tours are known to be less common. For both these reasons, a slightly different algorithm was used for removing intermediate stops from the ATRI data. This algorithm focuses on truck trips longer than 50 miles. As shown in Figure 15, if a truck traveled more than 50 miles from point A to point B, and then from point B to point C (points B and C are assumed as potential stops based on dwell time), then the distance between A and B plus the distance between B and C was compared to the direct distance between A and C. If the trip length of the direct path from point A to point C was longer than 95% of the sum of trip length from point A to point B and from point B to point C, then point B was categorized as an intermediate stop and removed. This method assumes trucks do not go far out of their way to fill the tank or for the driver to rest, which is generally a reasonable assumption. This approach filters out many intermediate stops while retaining actual stops on long-distance tours.

Figure 15: Reprocessing of ATRI data to Filter Out Intermediate Stops on Long-Distance Truck Trips



The resulting truck trips were once again subtracted from the total cell phone-based trips after reprocessing the ATRI data using the algorithm to remove intermediate stops. The ideal scenario is zero OD pairs with negative trips after subtracting the truck trips from the total trips; more generally, fewer negative OD pairs means a better reconciliation and fusion of the data. The resulting trip table still included some negative cells due to inconsistencies in the methods of filtering intermediate stops in the two datasets and general sampling errors. However, number of OD pairs with more truck trips than total trips was reduced by 87%—from nearly 11% of the ODs to only 1.4% and involving less than 0.1% of the total trips. (See Table 19.) Although still not perfect, the output was deemed acceptable because more than 98% of all cells have reasonable auto trips comprising 99.9% of total trips. The remaining OD pairs with more truck trips than total trips were reduced to a fraction of a trip (to retain the information that some sort of trip was observed and allow for expansion given some possibility that a passenger trip may have been

observed). Table 19 reports the statistics of the long-distance passenger car trip table obtained by subtracting ATRI truck trips from AirSage total trips per the aforementioned methodology.

Table 19: Auto Trip Statistics After Combining AirSage and ATRI

Statistics	Unfiltered ATRI	Filtered ATRI
Total Trips from AirSage	43,112,009	43,112,009
Number of AirSage ODs	370,108	370,108
Truck Trips from ATRI	238,932	405,550
Number of Negative ODs	39,637	5,071
Total Negative Auto Trips	99,049	39,303

Figure 16 and Figure 17 show the reduction in negative auto trips by comparing problematic origins after dropping unfiltered and filtered ATRI from AirSage, respectively. As seen in these figures, many of negative auto trips are along major interstates and long-distance routes, which could be due to unfiltered intermediate stops. Other remaining problems are boundary effects where trips are truncated as they enter or leave the model area. These figures confirm the statistics reported in Table 19 that there are fewer negative auto trips when filtered ATRI trips are subtracted from AirSage. The ATRI filtering process improves the auto trips significantly and generates more reasonable and more accurate auto trip tables. The remaining negative auto trips indicate the likelihood of some remaining issues with the intermediate stops (or perhaps coverage drops along interstates), but the magnitude of remaining issues is greatly reduced.

Figure 16: Negative Auto Trip Production with Unfiltered ATRI Trips

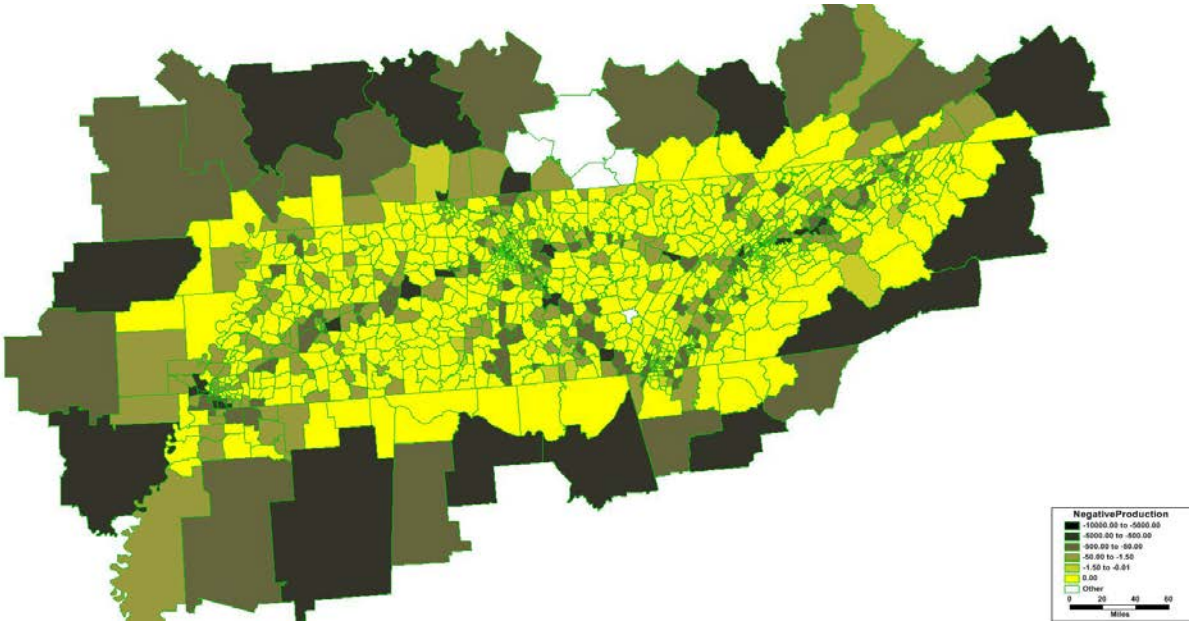
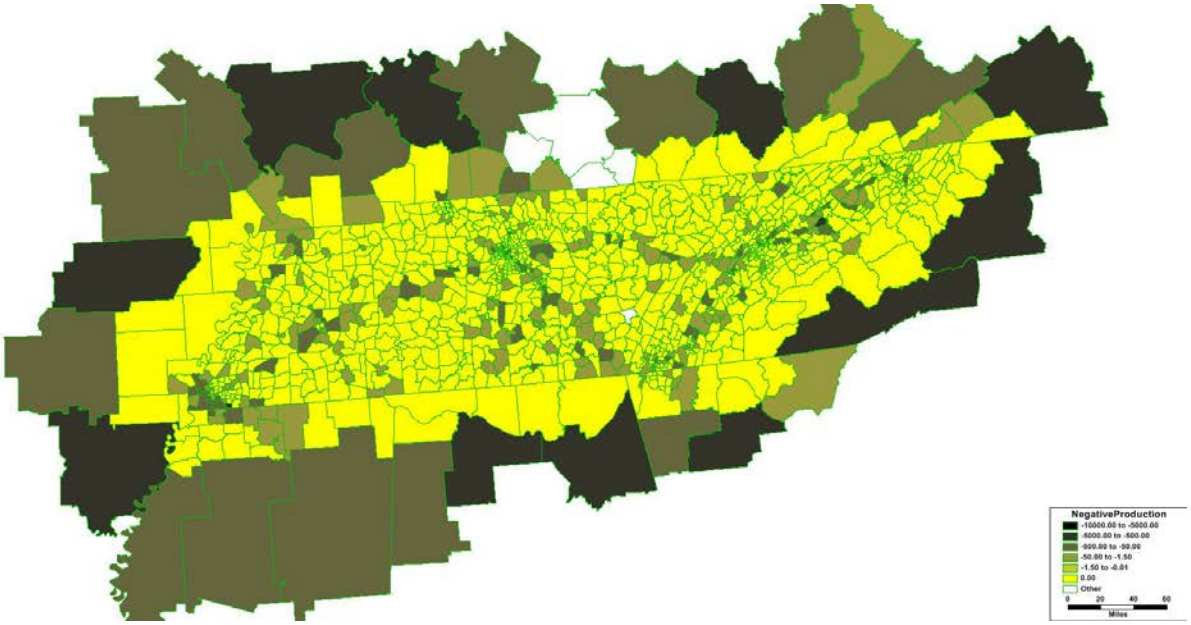


Figure 17: Negative Auto Trip Production with Filtered ATRI Trips



4.0 Reconciling Passive OD Data and Traffic Counts

Contemporary data-driven highway forecasting relies jointly on passive OD data and traffic counts. For several reasons, neither data type is sufficient for all forecasting applications, but together these two data resources can help provide robust answers to critical forecasting questions. This section presents available methods for reconciling and combining these two very different types of data, with illustrations from several applications.

4.1 The Need to Combine Traffic Counts and Passive OD Data

Neither traffic counts nor passive OD data alone can support forecasts for many practical highway forecasting applications. Passive OD data cannot be used with confidence unless it can be expanded to accurately represent all travel of interest. Traffic counts alone cannot provide forecasts in any scenario where diversion is likely to occur.

Consider, for instance, the need to forecast demand for proposed new managed lanes on an existing freeway. The likely traffic and revenue are importantly related to the current traffic counted on the freeway, but the current traffic count provides no information about how traffic patterns might change in response to the new lanes. The amount of traffic that might divert to the new facility to take advantage of the additional capacity and faster travel provided by the new lanes is critically related to the number of trips between origins and destinations served by the route. Passive OD data can provide an understanding of the number of trips. However, it would be naïve to accept this number as correct if it does not reconcile with the traffic currently counted on the facility and other parallel or related facilities.

To offer another example, forecasts may be required soon to help evaluate policies such as limitations on deadheading by autonomous vehicles or zero-occupant vehicle (ZOV) trips, such as restricting these trips during peak hours. Current traffic counts are of little direct help in such a forecasting application. A reasonable forecast of potential future ZOV trips must begin with an accurate understanding of where occupants are currently taking vehicles. Passive OD data can provide this information, but only after it has been properly expanded to and reconciled with traffic counts.

As the previous section explains, all existing commercially available passively collected OD data are based on incomplete sample frames. These commercially available datasets exclude travelers without mobile devices while they travel, and these datasets include only a select portion of travelers with mobile devices. Moreover, short-distance trips or short-duration activities are often under-represented in the data because they require more frequent observations of position. Travel to and from locations with poor coverage can also go un- or under-detected. Failure to account for such biases can lead to erroneous representations and faulty predictions of trip lengths, trip flows between origins and destinations, and present and future travel activity and traffic in general.

Traffic counts provide unbiased information on the spatial distribution of traffic. Traffic counts are currently the only data available to support expansion methods for passive OD data capable of correcting systematic biases related to coverage (rather than market penetration) and trip length or activity duration. The following section provides more information and illustrations of various methods by which traffic counts can be used to expand OD data.

Some analysts or modelers are reluctant to "mix" supply and demand data in this way. This reluctance may be rooted in the idea that traffic counts should provide independent validation of demand estimates developed solely from other sources. The development of travel demand models is sometimes presented in this way, but this is extremely misleading. In actual practice, demand estimates, whether based on "pure" synthetic models or directly observed data, are always adjusted to reconcile with or "validate to" traffic counts. The acknowledgement of this and the use of traffic counts in a well-defined process of expanding or adjusting demand estimates should be preferable to their use to adjust demand estimates in a series of ad hoc and often poorly documented manual adjustments. Ideally, validation would be done using before and after data and testing the ability of the model to replicate observed changes, but in cases where this data is unavailable and concern about the lack of independent validation remains, careful choice and use of expansion methods can address this by leaving a holdout sample of counts that are not used in the expansion process. These counts can provide independent data validation. Matrix partitioning or iterative screenline fitting methods are particularly amenable to this and are discussed along with other methods in the following section.

4.2 Methods for Expanding and Reconciling Passive OD Data with Traffic Counts

Multiple methods exist for expanding passively collected OD data. A taxonomy of expansion methods, focused on methods used in practice and particularly on methods using traffic counts, is presented in the following section, which discusses and illustrates the various methods. In all, seven methods are presented, divided into various categories. At the highest level, these methods can be divided into two categories depending on whether they use traffic counts or whether they rely on another estimate of sample penetration. Two methods of expansion that do not use traffic counts are discussed because they are in common use. The remaining five methods use traffic counts in various ways to expand passive OD data. These can be divided first based on whether they use only a single or multiple expansion factors. The latter can then be divided based on whether they make use of a network assignment model. Those that do make use of a network assignment model can be further divided into parametric methods that do not use ODME algorithms and nonparametric methods that do. Finally, ODME-based methods can be divided into those that rely directly on an ODME algorithm and those that use ODME to develop a simplified set of expansion factors.

The seven methods presented here have been included because they are known to be in use in practice, although level of use varies considerably. On the one hand, this volume does not endorse a single method, but on the other hand, it is not accurate to characterize all methods as equally robust or appropriate. The following discussion highlights both the advantages and limitations of various methods, recognizing that different methods or combination of methods may be appropriate for different applications. In fact, the authors have used six of these seven methods in different contexts, often in combination. The use of multiple methods in combination, while adding complexity and some measure of effort in some cases, can help address the limitations of one method with a second.

Figure 18: Taxonomy of Practical Methods of Passive OD Data Expansion

- Other Sample Penetration Methods
  - **Market Penetration**
  - **Trip Generation-Based**
- Traffic Count Methods
  - **Simple Scaling**
  - Variable Scaling
    - Matrix Partitioning
      - **Iterative Screenline Fitting**
    - Network Assignment-Based
      - **Parametric**
      - Nonparametric (ODME)
        - **Direct ODME**
        - **Indirect ODME**

4.2.1 Market Penetration-based Methods

Market penetration-based methods are one of the most basic approaches to expanding passive OD data, and the most discussed in academic literature (16,17,18,19). This method involves expanding the data by dividing by an estimate of the sample penetration rate, perhaps varying by location.

This approach is used by AirSage in expanding their data drawn from cell phone service providers. The number of subscribers to their services in each Census tract or Census Block Group are known to AirSage, as is the total population in each tract or block group. The ratio of subscribers to the total population provides their market penetration in each area. AirSage can assign each device to a home tract or block group with a reasonable degree of confidence since they have access to persistent device IDs—although anonymized—based on observations over one month or more. They can then expand the observations for each device by dividing by the market penetration in its home tract or block group. There may be further adjustments that AirSage makes as part of their expansion, but this is the basis of their method and any further adjustments are proprietary trade secrets.

This approach has several advantages. This approach can help address and correct for variations in sample penetration related to cell phone service providers' market shares, which are known to vary both across and within regions. Moreover, to the extent that tract or block group residents are relatively homogenous, this approach can help correct demographic biases. For instance, if a cell phone service provider has a higher market share in a higher-income or elderly neighborhood, observations from this subsample will be factored down; whereas, if the provider has a lower market share in a lower-income or younger neighborhood, this subsample will be factored up. In this example, the unexpanded data would be biased toward the affluent and elderly, but the expansion would at least help reduce this.

However, this approach also has several important disadvantages. It has been used with other datasets in research, but is not believed to be widely used in practice except by AirSage. A key obstacle to this approach is that many datasets do not have the kind of market penetration information that AirSage has access to for its adjustments. In addition, for cell-based data, it is unclear whether subscribers who receive a cell phone from their employer are reported at their

place of work or residence (or inconsistently) and this (among other factors) can skew the calculation of market penetration and make it inconsistent with the imputation of residence location within the dataset. More importantly, this method cannot correct systematic biases in the data arising from poor coverage (in areas traveled to as opposed to resided in, as this would presumably affect market share) or arising from the infrequency or inconsistent frequency of observations and resulting in higher rates of detection of longer trips.

Given these last two considerations, use of this method alone risks significant systematic biases in the data. However, there is little reason not to combine this method with another that can remedy these issues. It is the standard practice of the authors to use AirSage's expansion as the first step in a multistep expansion process that also involves the use of traffic counts to produce final expansion factors for AirSage data.

4.2.2 Trip Generation-based Methods

A similar method which does not require market share information instead estimates the sample penetration rate in travel analysis zones by comparing the number of observed trips with the number of trips predicted by trip generation models. Preliminary expansion factors are then taken as the inverse of the estimated sample penetration as in the previous method. An expansion factor can be estimated for both a trip's origin and its destination, so final expansion factors are produced using iterative proportional fitting on the trip OD matrix.

This method does have some advantages. This approach does not rely on proprietary market share information and can be applied to any dataset. Moreover, the approach has been shown to reduce the error in the data when compared to counts using a network assignment model. (14) This method can also address systematic location-specific biases arising from "holes" in coverage (provided there is at least enough coverage to observe a small number of trips to or from the area). However, as with the previous method, this approach is generally unable to address systematic biases related to frequency of observations and trip length.

However, the approach has some disadvantages as well. First, it relies on trip generation models, which can themselves have considerable error. Trip production models based on local survey data can produce reliable estimates of home-based trip-making, but trip attraction models and non-home-based trip generation models can introduce considerable error and provide a questionable basis for expanding passive OD data. This is of particular concern in areas where trip generation models are not based on local survey data; it is difficult to support the use of national defaults or averages to adjust actual observations of the local area from passive OD data.

Given these last considerations, use of this method alone risks significant systematic biases in the data. However, it could be tried in combination with other methods that use traffic count data.

4.2.3 Simple Scaling to Traffic Counts

Simple scaling to traffic counts produces a single estimate of global sample penetration for the data by comparing the passive data to traffic counts at one or more locations. This method is believed to be common in practice, although it difficult to document as it is rarely presented or published due to its simplicity. In the case of GPS-based data, traffic counts may be directly compared to observations of vehicles on a facility. In other cases, traffic counts can be compared

to the passive OD data by assigning the trips to roadway facilities using a network assignment model that can introduce its own error. Streetlight Data now offers a webtool for doing this type of expansion with its GPS data.

Sometimes, as when this method is applied as a second step after a market penetration-based expansion, the resulting scaling or expansion factor is explained as correcting for automobile occupancy or the number of mobile devices per vehicle. However, this is only one of several factors that may be captured by and reflected in this scaling.

The simplicity of this method is clearly an attractive feature as it is easy to explain to nontechnical audiences. Moreover, it requires little effort to apply the method, and traffic counts help guarantee a high level of consistency in traffic observed in the traffic counts and in the expanded passive OD data.

However, as with the similarly simple market penetration-based method, this approach is not capable of correcting systematic biases in trip length or related to holes in coverage, and therefore, similarly, use of this method alone risks significant systematic biases in the data. Also, like the market penetration-based method, this simple scaling is often used as the first stage in a multistep expansion process.

4.2.4 Matrix Partitioning/Iterative Screenline Fitting

Matrix partitioning/iterative screenline fitting is unique in that makes use of traffic counts to produce several expansion factors that may correct for systematic biases without using a network assignment model. Avoiding the use of a network assignment model is an advantage since the use of any model can introduce error. Moreover, this approach typically can only make use of a subset of traffic counts in a region, resulting in a holdout sample of counts that can still be used to provide independent validation of the passive OD data.

The approach works by first identifying “screenlines” or “cutlines” such as are commonly used to validate travel models. Each screenline should partition the study region into two subareas and align with the zone system used to define ODs, and traffic counts should be available or taken everywhere the roadway network crosses the screenline. (For this reason, it is helpful to choose screenlines that follow natural/physical barriers such as rivers, freeways, and railroads that have limited roadway crossings.) The sum of the traffic counts along each screenline can then be compared to the number of trips in the OD matrix that cross the screenline.

This comparison can be made with a network assignment model by partitioning or aggregating the OD matrix. Since each screenline partitions the region into two subareas, A and B, all origins and destinations can be identified as falling in either A or B. The OD matrix can then be aggregated into a matrix of just four cells: trips from A to A, trips from A to B, trips from B to B, and trips from B to A. The two off-diagonal cells (trips from A to B and from B to A) cross the screenline while the others do not. In this way, groups of OD trips can be compared against screenline counts without a network assignment model, and a preliminary expansion factor developed as the sum of the screenline counts divided by the sum of the off-diagonal elements of the aggregated matrix. The iterative screenline fitting process works by iterating or looping over the screenlines, factoring trips crossing each screenline to match the screenline counts. Although this factoring guarantees that the OD trips match the sum of counts for the current screenline,

each factor can introduce disagreement between the OD trips and previous screenlines. This is because individual OD pairs may cross several screenlines and have several differing factors applied. For this reason, the iteration is needed so that the expansion factors for individual OD pairs can stabilize to values that minimize errors versus all the screenline counts (but do not guarantee perfect agreement of the OD data with any individual screenline count).

This approach is not believed to be widespread; however, the authors have seen this method applied in several regions, including Anchorage, Alaska; Chattanooga, Tennessee; and San Diego, California. The relative value of the approach compared to simple scaling to total counts and its ability to address systematic biases in the passive data is largely a function of the number of screenlines that can be constructed for use in the procedure. A moderately large number of screenlines may be required to fully correct for trip-length-related biases as ODME-based methods can and it may be difficult to construct many screenlines in some areas.

However, ironically, it can also be of use in areas with a relatively limited number of counts that are not suitable to ODME if at least a small number of screenlines can be constructed. Although it will only produce incrementally more value than simple scaling to total counts in such cases, this increment of value may be important. For example, San Diego had limited truck counts that were not well-suited to ODME-based methods, but screenlines could be constructed which basically created a cordon around the metropolitan area that allowed importantly differential expansion of trips to and from the metro area versus trips within the metropolitan area.

Figure 19: Screenlines for Matrix Partitioning in Chattanooga

The Chattanooga application (13) provides a contrasting example where at least a moderately large number of screenlines (shown in Figure 19) could be constructed owing to the mountainous topography and many waterways that limit street connectivity along several ridgelines and waterways crossing the region. Compared to assignment of OD data expanded using simple scaling, assignment of the Chattanooga OD data expanded using this method yielded substantially improved agreement with traffic counts overall, not just the traffic counts along the screenlines used in the process. This results provided some validation of the process and its viability as an alternative to ODME-based methods where enough screenlines can be constructed.

4.2.5 Parametric Scaling to Traffic Counts

Parametric scaling to traffic counts is perhaps the most straightforward way of addressing and correcting for systematic trip-length biases in passive OD data. In theory, this method may be able to be applied to GPS-based datasets without the use of a network assignment model, but this would require substantial data processing and, in practice to date, it is only known to have been applied using a network assignment model. The approach can also theoretically address some coverage issues, although this also remains untested in practice as of the writing of this volume.

The approach is to estimate the parameters of a formula that produces expansion factors for trips as a function of their length or other attributes. The parameters are estimated using least squares error (LSE) versus traffic counts. The parameter estimation can be formulated as a bilevel programming problem, but it is particularly difficult (NP-Hard) as it involves an equilibrium constraint in the lower level traffic assignment problem. Hence, metaheuristics (e.g., genetic algorithms) are typically used to solve for the parameters. (Although simpler line search methods can be used if the expansion factor is modeled as a simple linear function of trip length, this approach is not recommended as evidence points to a nonlinear relationship and the significance of other factors.)

This approach has the advantage of producing relatively easily understood expansion factor formulas and avoiding the ambiguities of ODME-based approaches. Moreover, it is firmly grounded in a robust statistical procedure, and can therefore, in theory for instance, be used to determine the statistical significance of systematic biases in the data. However, the involvement of a network assignment model and resulting need to employ metaheuristics to estimate the parameters of the expansion factor function make the process both mathematically complex and computationally intensive. As a result, the approach may not be practical for some practitioners without access to the appropriate tools or background knowledge.

The development of the Tennessee statewide model used this method as the third stage of a four-step approach to data expansion. (14) The overall approach began by taking the market penetration-based expansion provided by AirSage and adjusting it separately for automobiles and trucks (after the data was partitioned using the ATRI truck GPS data) using simple scaling to total auto and truck counts. However, these first two steps still resulted in poor agreement between the OD data and traffic counts, with substantial underloading in urban areas on the order of -10% versus counts and substantial overloading in rural areas on the order of +15% versus counts.

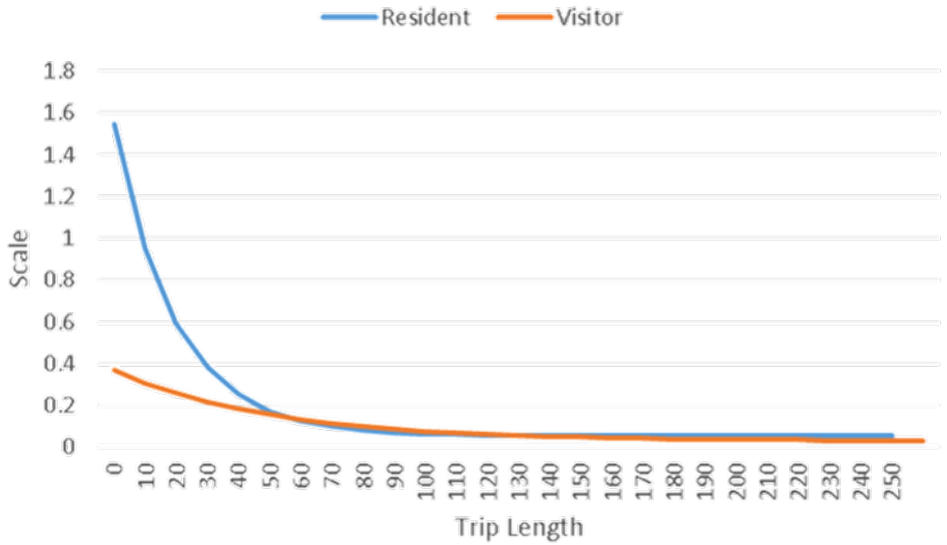
Parametric scaling was then applied, followed by direct ODME adjustments. These final two steps resulted in substantial improvements in the agreement between the OD data and traffic counts. Moreover, this combined approach helped avoid large adjustments from ODME without a clear understanding of the underlying problem or issue, and clearly confirmed the substantial, systematic trip-length bias in the data as well as how it differed between Tennessee residents and visitors in the state. This approach also helped demonstrate and confirm that trip-length bias is not an artifact of the tendency of some ODME algorithms to use short-distance trips to artificially improve the fit of OD data to specific counts.

Expansion factor functions were estimated separately for resident and visitor trips. This was done because visitor trips exhibited more consistent over-representation independent of trip distance. This is consistent with the general hypothesis of a bias against short trips since in this context visitors are already, by definition, long-distance travelers. A genetic algorithm was used to fit least squares parameters for the function: expansion factor = c + a x Exp (b x distance). The curves can be seen in Figure 20.

For resident trips, c = 0.0612, a = 1.6404, and b = -0.0507.

For visitors, c = 0.0292, a = 0.3376, and b = 0.0195).

Figure 20: Distance-based Expansion Factor Functions for Resident and Visitor Trips in Tennessee



The implication of the resident curve is that a 100-mile trip is 12 times as likely to be detected in the cell phone data as a 10-mile trip. Given that there may be two to three times as many people on a 100-mile trip as a 10-mile trip, this suggests that a 100-mile trip is four to six times as likely to be detected than a 10-mile trip for reasons other than vehicle occupancy. The application of these scaling factors did not completely resolve the observed loading errors, but it significantly improved them by reducing urban underloading to approximately -2% and rural overloading to approximately 5%.

In sum, the parametric scaling method is a powerful method for understanding and addressing systematic biases. This method can be used in combination with ODME to limit the need for and magnitude of ODME adjustments, but its mathematical complexity and computational intensity may not be practical for all applications.

4.2.6 Direct ODME

The direct use of ODME algorithms to expand passive data to traffic counts is believed to be one of the most common approaches in practice and is also widely documented in the literature. (5,7,8,15,20,21,22) It is important to recognize that there are several different ODME algorithms in use and that different algorithms can produce significantly different results and have different properties. (A good review of ODME techniques, their limits, and effectiveness can be found in the study by Marzano et al. [23]) ODME methods, which use OD data only as a "seed" or starting point and produce a final adjusted OD matrix purely by minimizing errors versus traffic counts, are not appropriate for expanding passive OD data as they can significantly distort the observed data. However, methods which attempt find a solution and produce a final OD matrix which minimizes errors versus counts and versus the original OD data or only with appropriate constraints on adjustments to the original OD data can be powerful and appropriate methods for data expansion. These methods are capable of correcting systematic biases related to trip lengths as well as coverage "holes" (provided there are at least some observations in the "holes" to expand).

A proper understanding of ODME is grounded in two important facts. First, counts do provide real information about underlying OD patterns. Second, counts alone cannot be used to identify OD patterns. Both facts can be proven mathematically. The truth of the former is demonstrated via the method of iterative screenline fitting. The truth of the latter is evident from the fact that the number of "known" traffic counts is always substantially smaller than the number of "unknown" OD flows; the problem is statistically under-determined and there is not a unique set of OD flows that correspond to a set of traffic counts on a network.

Regarding the first fact, since counts provide information about the underlying OD patterns, ODME has the potential to improve or correct OD matrices from Big Data. Regarding the second fact, since counts alone cannot identify OD patterns, ODME methods focused solely on count data are ill-conceived. A balanced ODME approach recognizes the value of both traffic count data and passive OD data and uses traffic counts to improve the representativeness of OD data while not distorting it. In fact, it is important to understand that mathematically, because the OD solution space dwarfs the network solution space, OD data is more important than count data in producing a good final solution. So long as an ODME method is used in a manner consistent with this fact, it can be an efficient and powerful tool for expanding passive OD data.

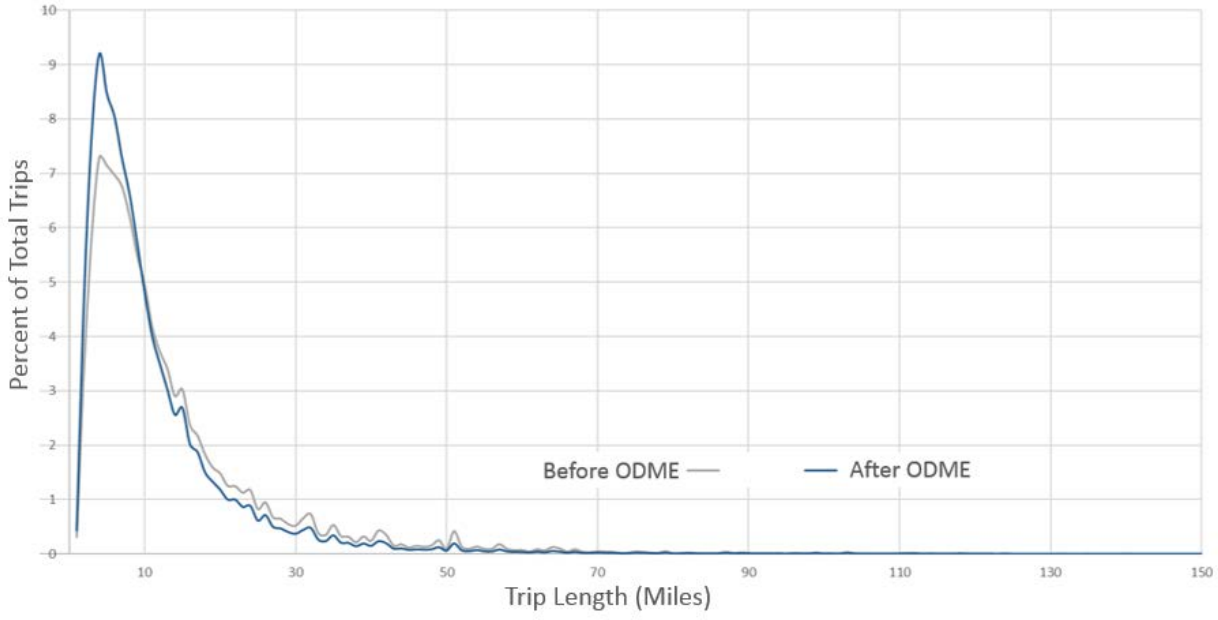
Direct ODME has several practical advantages as a method to expand passive OD data. Since software implementations of ODME algorithms are widely available, direct ODME is one of the quickest and easiest ways of expanding OD data to traffic counts and correcting for systematic trip-length biases. In fact, ODME can correct for a variety of different types of errors or biases in the OD data without requiring complex methods or in-depth analysis. However, ODME can over-correct and distort OD patterns to over-fit count data if not used carefully and appropriately. This danger and the distrust that it inspires in some professionals is the main drawback of the method

together with its lack of transparency and the difficulty of understanding the underlying issues that the expansion adjustments are addressing.

As noted in the previous section, direct ODME was used as the final step in a multistep process to expand the passive OD data for Tennessee statewide model. ODME was applied last so that it could be used in a more limited way to address only the issues that other methods could not resolve or could not address as well. Careful consideration was given to setting appropriate bounds on the ODME adjustments. On the one hand, the most limited adjustments capable of producing good agreement with counts are desirable. At the same time, it is important to acknowledge and allow ODME to factor trips to and from certain areas up and down to account for varying degrees of cell coverage and other factors that can cause necessary expansion factors to vary beyond simply the variance in cell phone market shares by resident areas. After some experimentation, a minimum factor of 0.5 and a maximum factor of 5.0 were chosen to limit ODME scaling of any given OD pair.

In addition to these limits, the average amount change in the trip matrix from ODME was closely monitored. The average absolute difference between cells in the final adjusted trip matrix and in the scaled matrix was 4.3 trips and the average absolute percentage difference was 1.5%. These were deemed to be generally reasonable adjustments together with the limits on minimum and maximum adjustments. The trip-length frequency distribution of the adjusted matrix was also compared to the original matrix (Figure 21). The comparison showed that ODME resulted in a modest additional increase in the expansion of short-distance trips versus longer trips. This seemed to suggest that the distance-based scaling was not excessive and was successful in accounting for most of the distance-related adjustments. The ODME adjustments improved the fit of the cell phone based data from 55.5% RMSE to 36.6% RMSE versus over 12,000 traffic counts across the State of Tennessee. This was deemed a successful and helpful improvement to the expansion given the relatively limited adjustments necessary to achieve this improved fit.

Figure 21: Tennessee Trip-Length Frequency Distributions Before and After ODME



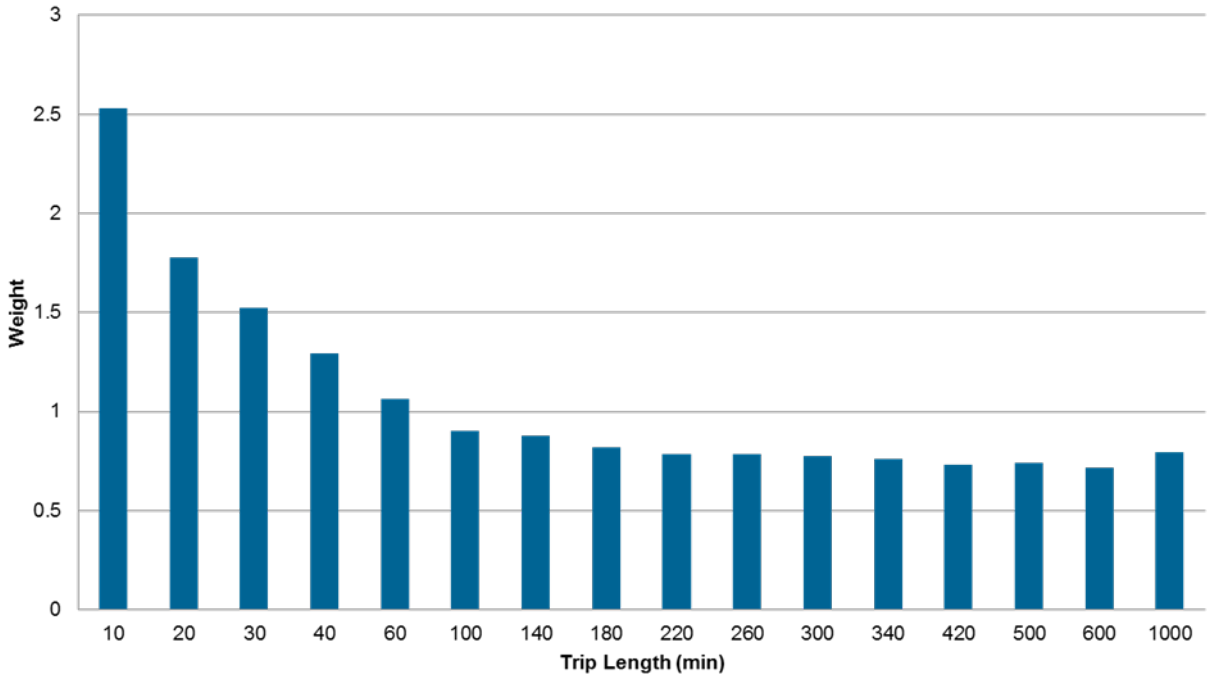
The Tennessee application provides a good illustration of how direct ODME can be used in a responsible and carefully controlled way to produce better agreement between passive OD data and traffic counts while ensuring the integrity of OD patterns observed in the Big Data.

4.2.7 Indirect ODME

Rather than using ODME directly to expand passive OD data, ODME can be performed and its results analyzed to produce a limited set of expansion factors. A more limited set of expansion factors can be more readily understood and interpreted than a multitude of direct ODME-based expansion factors, inspiring greater confidence in some cases. Moreover, this approach can help establish the amount of the ODME adjustments related to phenomena such as trip length. It can also confirm that these adjustments, such as changes in average trip lengths in themselves result in better agreement between the OD data and traffic counts independent of the details of the ODME adjustments.

The main advantages of this approach are its relatively high level of transparency and interpretability of results compared to ODME, its support of insights from ODME, and its modest level of effort and leveraging of widely available ODME algorithms. The level of effort associated with the approach can vary depending on the complexity of the expansion factors developed. Basic schemes to address trip-length bias can be applied with only marginal additional effort compared to direct ODME, while complex schemes using multiple factors can require substantial effort. The additional increment of effort beyond direct ODME is one of the disadvantages of the approach. This method also cannot produce as good agreement with counts as direct ODME or correct for errors in the data that are more difficult to understand or identify.

Figure 22: Expansion Factor Scheme using Trip Length for Iowa Truck GPS Data



The indirect ODME approach was used to expand truck GPS data for application in the Iowa statewide model. (7) A simple scheme of expansion factors based on trip length (duration) was

used and the resulting factors are shown in Figure 22. Consistent with systematic trip-length bias, short-haul truck trips (<60 minutes) were factored up while long-haul trips (>60 minutes) were slightly factored down. The resulting expansion factors ranged from slightly over 0.7 to just over 2.5; the latter value only applying to truck trips under 10 minutes in duration. Trips that were 10–20 minutes in duration only had to be factored up by less than 1.8. This does indicate the need to correct the data for this bias, but it also suggests that a simple and reasonable weighting scheme can produce reasonably good results.

When assigned to the highway network, truck GPS data simply scaled to total counts resulted in a RMSE of 116% versus an RMSE of 92% after applying the indirect ODME-based factors to the data or 58% RMSE resulting from direct ODME. Thus, the simple trip-length-based expansion factors from the indirect ODME approach accomplished over 41% of the improvement in fit to the counts provided by direct ODME. This both confirmed that the trip-length bias was not an artifact of ODME and suggests that ODME is likely also correcting other errors or biases in the data, although it is also possible that the ODME may be overfitting the count data. This example shows how indirect ODME can add value and support insight beyond direct ODME methods; however, determining whether this is worth the additional effort may depend on the application.

5.0 Data-Driven Traffic Forecasting and Modeling

The preceding sections have explained how to process traffic counts and passive OD data and combine them to produce the best possible estimates of OD matrices describing existing travel patterns. However, OD matrices are not produced for their own sake, but as means of producing traffic forecasts. For many practical short-term forecasting applications, OD matrices themselves are enough to produce forecasts together with some form of well-calibrated network assignment or traffic simulation model capable of predicting rerouting in response to changes in the highway network. Since network modeling methods and standards are well documented elsewhere, readers interested in this sort of forecasting application may choose to stop reading here.

However, in some cases, when OD travel patterns may change in response to network changes or land development or use, network modeling and OD matrices are not enough by themselves to produce reasonable forecasts. In these applications, common both for long-range planning and for major environmental studies that require the consideration of secondary and cumulative impacts (such as changes in land development and use), forecasts require not only a firm understanding of existing OD travel patterns, but also an understanding of how they might be expected to change. Predicting changes in OD travel patterns requires travel demand modeling in addition to travel demand data. Travel demand models have a long and well-established history predating the more recent development of OD matrices from passive Big Data. Thus, it is generally well understood how travel models can be used to produce forecasts of future OD travel patterns in the absence of Big Data, but in the current context it is important to consider how travel models can produce forecasts of OD patterns with Big Data. In other words, how can travel demand models be used together with the type of Big Data-derived OD matrices developed in the preceding sections of this volume?

There are generally two methods for using travel demand models together with passive OD data or incorporating passive OD data in travel demand models. The first approach uses travel demand models (usually of more traditional, aggregate designs) to pivot off OD matrices developed from Big Data and traffic counts. The second approach instead uses these OD matrices to develop fixed factors (or constants) that are incorporated into the travel model; this approach is more attractive for activity-based demand simulation models, although it can also be applied with aggregate trip-based travel models. The following sections describe and discuss these two similar and related, but alternative and potentially different, approaches.

5.1 Pivot-Point Methods

The most common approach to using travel demand models together with an independently data-derived trip matrix is to apply the change in OD travel patterns predicted by a model to the data-driven OD matrix. (24,25)

This approach often uses rules or a weighting scheme to combine additive pivoting and multiplicative pivoting, but pure additive pivoting is also sometimes used. Pure multiplicative pivoting is to be avoided because it can result in unreasonable results in cells where the synthetic model and actual data differ significantly in the base case.

5.1.1 Additive Pivoting

Additive pivoting predicts the alternative demand (Â) by subtracting the modeled OD matrix for the base case (B_M) from the modeled OD matrix for the alternative (A_M) and adding this difference to the data-derived OD matrix (B_D).

Equation 12: Additive Pivoting

$$\hat{A} = B_D + (A_M - B_M)$$

5.1.2 Multiplicative Pivoting

Multiplicative pivoting edicts the alternative demand (Â) by dividing the modeled OD matrix for the alternative (A_M) by the modeled OD matrix for the base case (B_M) and multiplying the data-derived OD matrix (B_D) by this growth factor.

Equation 13: Multiplicative Pivoting

$$\hat{A} = B_D \times (A_M \div B_M)$$

Multiplicative pivoting is sometimes preferred for normal, moderate growth or changes, but can produce poor forecasts in some cases, particularly when there are very few or no trips for an OD pair in one or more of the matrices.

5.1.3 Composite Pivoting

Rules or weighting are commonly used to select or combine the two basic pivoting methods, additive and multiplicative. The best documented of these combined, rule-based approaches is the “eight-case” method; (24,25) however, this method requires the choice of a breakpoint for distinguishing between normal and extreme growth and switching methods. The selection of this breakpoint can require experimentation and may not be generalizable between different types of studies, etc. Therefore, some models (e.g., the Indiana Statewide Travel Demand Model) use simple additive pivoting. Theoretically, approaches based on averaging the additive and multiplicative methods (such as is used for pivoting modeled roadway volumes off of traffic counts in the Ohio Department of Transportation’s certified traffic procedures) could also be used, but examples of this type of approach for pivoting OD demand are unknown to the authors.

5.1.4 Advantages and Limitations of Pivoting Methods

Pivot-point methods have the clear advantage of requiring relatively little or no modification to an existing travel demand model and hence relatively little effort to apply for an individual study. However, when incorporated within a model rather than used for an individual forecast, they can require careful management and updating of an input file for the base-case modeled OD matrix which must be kept current with the model’s calibration. This has little impact on the application of the model for routine forecasting, but it can complicate updates to the model, including zone splits.

Pivot-point methods also are attractive because they are straightforward and easy to understand in concept and explain. Many professionals are already familiar with pivot-point methods from their use to pivot off individual traffic counts to produce facility-specific forecasts. Pivoting on ODs rather than highway network link volumes is less familiar to many in the United States, but it has

long been common in Europe and Australia and is quickly growing in use in the United States in response to the advent of passive OD data.

Pivot-point modeling can substantially improve forecasts by removing the error in a travel demand model's base-case OD matrix. This error is known to be the largest source of error in traffic modeling (1); thus, pivot-point methods promise substantially improved accuracy in forecasting. However, pivot-point methods have no effect on the sensitivity of the travel model or resulting forecast to changes in travel time, tolls, land use, or other factors. This can be viewed in either a positive or negative light. On the one hand, the independence of the model's sensitivity to the approach can alleviate any concerns related to overfitting or over-specification. On the other hand, this same independence of the model's sensitivity to the approach also means that the information in the passive OD data does not necessarily improve the sensitivity of the travel model or resulting forecast to changes in travel time, tolls, land use, or other factors. The large amount of error in base-case models suggests the strong possibility of under-specification errors in existing or traditional models which may translate into over-sensitivity of models to travel times, tolls, land-use variables, and other factors, and pivot-point methods do not help to address this issue.

While the inability of pivot-point methods to address under-specification errors affecting model sensitivities is an important theoretical concern, one of the main drawbacks of pivot-point approaches in practice is the inability of applying the approach at the level of disaggregate demand in demand simulation models such as activity-based models or supply chain simulation models. The fixed-factor approach presented in the following section offers an alternative method that can be applied to disaggregate demand simulation models as well as traditional aggregate models.

In summary, pivot-point approaches may not be theoretically ideal or practical for use with activity-based or supply chain simulation models, but they are easy to apply with many travel models and can substantially reduce error.

5.2 Fixed-Factor/Constant Rich Methods

Fixed-factor or constant rich approaches involve a deeper integration of passive OD data into a travel model. As such, they generally require more effort, but they can also potentially yield greater benefits than pivot-point methods and are applicable to activity-based or supply chain simulation models as well as more traditional aggregate trip-based models.

The fixed-factor approach works by incorporating a set of constants into the spatial (gravity, destination, or activity location choice) model components of a travel demand modeling system. These factors are estimated in a statistically rigorous way to allow the model to reproduce expanded passive OD data with minimal error.

Fixed factors or constants can be specific to individual or groups of origins or destinations or OD pairings. In the context of destination choice models, these are alternative specific bias constants.

Fixed-factor methods can be developed in two importantly different ways. First, a sequential estimation approach in which the factors are estimated after and independently of other model parameters is like pivot-point methods in that it does not affect model sensitivities for good or ill, and it is easier to apply. This method usually involves estimating the constants as shadow prices. Second, simultaneous estimation of fixed factors together with other model parameters requires

more effort, but it also offers the potential for better results by addressing likely under-specification errors and potential model over-sensitivities. Over-specification errors are still possible, though this is less of an issue with Big Data.

5.2.1 Shadow-Pricing

A constant rich, fixed-factor approach was used to incorporate passive OD data in the activity location choice model components of a DaySim activity-based modeling system for the Chattanooga, Tennessee, MPO. (13) In this application, believed to be the first in which passive OD data has been incorporated in an activity-based model, the simpler, sequential estimation approach was used and fixed factors were estimated using shadow-pricing techniques to minimize square error versus the OD data. The term shadow price is taken from economics where it is used for the Lagrange multiplier corresponding to the constraint on a demand function that the market reaches equilibrium. It is an unobserved "price," or factor, in the demand function that can be inferred from the observed point at which equilibrium is reached in a market. This corresponds well with this usage in travel modeling. In this context, the shadow price is an additional term in the utility function of the demand models that can be inferred from the actual observed travel patterns (in this case, from the passive OD data).

In a sequential, shadow-price approach, the deterministic utility (V_i') of a destination (i) is updated by adding a shadow price (s_i) to the originally estimated deterministic utility (V_i) which is held fixed.

Equation 14: Updated Utility with Shadow Price

$$V_i^{'} = V_i + s_i$$

The shadow price is estimated by iteratively adjusting it until application of the model reproduces the observed share of trips to the destination. This is typically done by setting the shadow price in iteration k+1 to sum of the shadow price in iteration k and the natural logarithm of the ratio of the observed share of trips to destination i (P_i) to the predicted share of trips to destination i in iteration k ($\widehat{P_i^k}$).

Equation 15: Iterative Updating of Shadow Prices

$$s_i^{k+1} = s_i^k + ln\left(\frac{P_i}{\widehat{P_i^k}}\right)$$

This sequential approach using iteratively developed shadow prices is relatively easy and a familiar procedure for modelers accustomed to working with activity-based models which typically use the technique to enforce the attraction or "double" constraint on work location choice.

The Chattanooga model's zones were grouped into 40 districts (in part due to limited confidence in the spatial precision of the data being used) and 1,600 constants were estimated for each district OD pairing. The shadow-pricing estimation was judged to have converged after 24 iterations. The sum of absolute errors versus the OD data decreased from 516,595 to 59,962 when comparing the original comparison of the model without constants to passive OD data to the final model with the fixed factors from shadow pricing. The weighted mean absolute percent error decreased from 22.2% to 8.3%, and the RMSE decreased from 37.1% to 10.5%. A summary

comparison of the actual patterns at the level of 12 further aggregated superdistricts is presented in Table 20 as a complement to these statistics.

Table 20 shows that a constant rich fixed-factor or shadow-pricing approach can allow travel demand models to reproduce travel patterns from passive OD data with good accuracy. The table shows that relative percentage errors for all OD pairs are between -0.4% and +0.6% and total trips from/to superdistricts are between -1.8% and +0.7% indicating very good agreement between the model and the data. The modeled OD matrices using these factors also produced improved agreement of the modeled traffic volumes with traffic counts on the highway network. It also contributed to good assignment validation statistics allowing the model to achieve a RMSE of modeled traffic volumes versus counts of 28.97% and a correlation coefficient of 0.971. Although some traffic counts were used to correct the expansion of the cell phone data using the iterative screenline fitting approach discussed in Section 4.2.4, the final agreement between the modeled roadway volumes and all traffic counts (including those not used in the screenline fitting) provides some degree of further independent validation of these methods and their ability to reproduce travel patterns.

Table 20: Comparison Between Passive Cell Phone Based ODs and Destination Choice Models with Fixed Factors in Chattanooga

ORIGIN SUPER-DISTRICT	DESTINATION SUPERDISTRICT												GRAND TOTAL
	1	2	3	4	5	6	7	8	9	10	11	12	
1	0.5%	0.1%	-0.2%	0.0%	0.0%	-0.1%	-0.2%	-0.2%	-0.1%	0.0%	-0.1%	-0.3%	-0.6%
2	0.2%	0.2%	0.1%	0.0%	0.1%	-0.1%	0.0%	0.1%	0.1%	0.0%	0.0%	-0.1%	0.5%
3	-0.2%	0.0%	0.3%	-0.2%	-0.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	-0.1%	-0.3%
4	0.0%	0.2%	-0.2%	0.1%	0.0%	-0.1%	0.0%	0.0%	0.0%	0.0%	0.1%	0.1%	0.3%
5	0.1%	0.1%	-0.1%	0.0%	0.4%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.7%
6	-0.2%	-0.1%	0.0%	-0.1%	0.0%	0.2%	0.2%	-0.1%	0.0%	0.0%	0.0%	0.0%	0.0%
7	-0.1%	0.0%	0.1%	0.0%	0.1%	0.0%	0.3%	0.1%	0.1%	0.0%	-0.1%	0.0%	0.5%
8	-0.1%	0.0%	0.0%	0.0%	0.0%	-0.1%	0.1%	0.3%	-0.1%	0.0%	0.0%	0.1%	0.2%
9	-0.1%	0.0%	0.0%	0.0%	0.0%	-0.1%	0.1%	0.0%	0.6%	0.0%	0.0%	0.0%	0.5%
10	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.2%	0.1%	0.0%	0.3%
11	-0.1%	-0.1%	0.0%	-0.1%	0.0%	0.0%	-0.1%	0.0%	0.0%	0.2%	0.1%	-0.2%	-0.2%
12	-0.3%	-0.4%	-0.2%	-0.2%	0.0%	-0.1%	-0.3%	0.0%	-0.1%	0.0%	-0.2%	0.0%	-1.8%
Grand Total	-0.1%	0.0%	-0.1%	-0.4%	0.5%	-0.3%	0.2%	0.2%	0.5%	0.4%	-0.3%	-0.7%	0.0%

5.2.2 Simultaneous Estimation

The alternative method to shadow-pricing is to estimate constants simultaneously with the rest of the utility function parameters. The final mathematical or functional form of the utility function is identical, but the estimates of the parameters may be different than those produced using shadow pricing. The maximum likelihood framework would be applied, but using a composite likelihood calculated against observations from both traditional household travel survey data and passive OD data. This approach allows the passive OD data to inform not only the pattern produced by

the model but also its sensitivity to travel times and other factors in the choice. This is theoretically superior to the sequential approach which begins with the estimation of the destination choice model without constants. It can be shown that, just as with regression models, estimating discrete choice models without constants (when constants are theoretically appropriate which they are in destination choice) can bias parameter estimates, skewing model sensitivities.

However, this simultaneous approach is more challenging in practice because it currently is not supported by any standard statistical or travel modeling software platform. It therefore requires the development of custom scripts for parameter estimation. There are two reasons existing statistical software do not support this procedure. First, it involves fitting parameters to two observed datasets (traditional household travel survey data and passive OD data) which remains an advanced topic. Second, in models with feedback from assignment to distribution, there is a lower level equilibrium constraint which greatly complicates the model. This and other considerations such as the inclusion of aggregate attraction constraints (which are subtly different than these constants even though they can both be similarly estimated and thought of as types of shadow prices) and/or the inclusion of accessibility variables in the model's utility function can make the likelihood function of the model nonconvex and require metaheuristics rather than the gradient based steepest ascent algorithms in common statistical software. For more information on the estimation of destination choice models, see the forthcoming TMIP How-to guide on that subject. In summary, while simultaneous estimation of fixed factors and other destination choice model parameters from both household survey and passive OD data (and possibly other data sources such as traffic counts) is a theoretically attractive, for the time being it remains a challenging approach in practice.

5.2.3 Differences from k Factors

It is worth noting that the constant rich methods presented above are importantly different than traditional k factors sometimes used in gravity models. The constants developed above are theoretically motivated, incorporated in a behavioral framework, and can be systematically statistically estimated from a sound support of passive OD data. In contrast, k factors were developed in an ad hoc fashion, with little or no theory, based on survey or traffic count data that often could not actually support them. Despite these important distinctions, some historical abuses of k factors still make some professionals hesitant or fearful of constant rich approaches. Individuals with a classical statistical background may also have a hesitancy due to fears of over-specification errors. However, while errors and abuse are possible in any statistical modeling, and some level of caution is always an important component in good judgment, in the new context of the availability of Big Data, conscientious professionals should reconsider constant rich approaches in an open and unbiased way. The emergence of a new generation of constant rich approaches is driven by a real change in the context of the data and analysis methods available. In addition to Big Data, machine learning analysis methods (such as the metaheuristics mentioned in the previous paragraph) are another new factor driving contemporary constant rich approaches that are also worthy of further consideration. Machine learning also provides a perspective that is more concerned with under-specification errors than over-specification errors, which may be helpful in balancing certain schools of classical statistical thought.

5.2.4 Advantages and Limitations of Constant Rich Methods

Constant rich approaches allow spatial choice models to incorporate passive OD data, better replicate observed OD patterns in the base case, and presumably better forecast future or alternative OD patterns. Moreover, constant rich methods can produce both agreement of aggregate OD patterns with observed data and consistency between the disaggregate and aggregate results of a simulation modeling system. In the context of simultaneous estimation of constants with other utility parameters, this approach should theoretically lead to less biased, more realistic model sensitivities, as well.

The main drawback to constant rich methods is the level of effort required, which is generally somewhat greater than the effort required for pivoting. However, the shadow-pricing method of sequential estimation of constants is only modestly more difficult than pivoting. Simultaneous estimation, despite its theoretical attractiveness, remains challenging in practice.

5.3 Conclusion

This volume has presented a broad set of methods for processing and combining traffic count and large-scale passive OD data and using them together with travel demand models to produce data-driven traffic forecasts. This topic has evolved rapidly over recent years and will likely continue to do so for some time. That said, this volume should remain a valuable resource for understanding practical methods and how they can be used to produce robust, data-driven highway forecasts. In turn, these forecasts may be used to provide accurate and relevant information to help answer critical questions about how transportation is changing or may change in the future.

References

1. Zhao, Y. and K. Kockelman. The Propagation of Uncertainty through Travel Demand Models: An Exploratory Analysis. *The Annals of Regional Science*. Vol. 36, No. 1, 2002, pp. 145-163.
2. *NCHRP Report 765: Analytical Travel Forecasting Approaches for Project-Level Planning and Design*, National Cooperative Highway Research Program, Transportation Research Board, 2014, NCHRP Report 765 link.
3. Gur, Y. J., S. Bekhor, C. Solomon and L. Kheifits. Intercity Person Trip Tables for Nationwide Transportation Planning in Israel Obtained from Massive Cell Phone Data. *Transportation Research Record: Journal of the Transportation Research Board*. No 2121, 2009, pp. 145-151.
4. Calabrese, F., G. Di Lorenzo, L. Liu and C. Ratti. Estimating Origin-Destination Flows using Mobile Phone Location Data. *IEEE Pervasive Computing*. Vol. 10, No. 4, 2011, pp. 36-44.
5. Bernardin, V. and L. Amar. Using Large Sample GPS Data to Develop an Improved Truck Trip Table for the Indiana Statewide Model. Presented at the 4th TRB Conference on Innovations in Travel Modeling, Tampa, FL, May 2012.
6. Huntsinger, L. F. and R. Donnelly. Reconciliation of Regional Travel Model and Passive Devise Tracking Data. Presented at the 93rd Annual Meeting of the Transportation Research Board, Washington, D.C., 2014.
7. Bernardin, V., S. Trevino and J. Short. Expanding Truck GPS-based Passive Origin-Destination Data in Iowa and Tennessee. Presented at the 94th Annual Meeting of the Transportation Research Board of the National Academies, Washington, D.C., January 2015.
8. Zanjani, A., A. Pinjari, M. Kamali, A. Thakur, J. Short, V. Mysore and S. Tabatabaee. Estimattion of Statewide Origin-Destination Truck Flows from Large Streams of GPS Data: Application for Florida Statewide Model. *Transportation Research Record: Journal of the Transportation Research Board*. No 2494, 2015, pp. 87-96.
9. Bindra, S., B. Grady and J. Deshaies. Using Cellphone Origin-Destination Data for Regional Travel Model Validation. Presented at the 15th National TRB Transportation Planning Applications Conference, Atlantic City, NJ, 2015.
10. Milone, R. Preliminary Evaluation of Cellular Origin-Destination Data as a Basis for Forecasting Non-Resident Travel. Presented at the 15th National TRB Transportation Planning Applications Conference, Atlantic City, NJ, 2015.
11. Zhang, W., A. Kuppam, V. Livshits and B. King. Evaluation of Cellular-based Travel Data – Experience from Phoenix Metropolitan Region. Presented at the 15th National TRB Transportation Planning Applications Conference, Atlantic City, NJ, 2015.
12. Lee, R. J., I. N. Sener and J. A. Mullins. An Evaluation of Emerging Data Collection Technologies for Travel Demand Modelling: from Research to Practice. *Transportation Letters: The International Journal of Transportation Research*. Vol. 8, No. 4, 2016, pp. 181-193.
13. Lee, Y., V. Bernardin and D. Kall, Big Data and Advanced Models on a Mid-Sized City's Budget: The Chattanooga Experience. Presented at the 15th National Tools of the Trade Conference, Charleston, SC, September, 2016.
14. Han, Y., K. Kaltenbach, S. Thomson, J. Balaji and D. Hulker. Innovative Analysis Methods of Mobile Phone Data in the Best Travel Demand Modeling Practice in Kentucky. Presented at the 15th National Tools of the Trade Conference, Charleston, SC, September, 2016.
15. Bernardin, V., N. Ferdous, H. Sadrsadat, S. Trevino and C. Chen. Integration of the National Long Distance Passenger Travel Model with the Tennessee Statewide Model and Calibration to Big Data. Forthcoming in *Transportation Research Record: Journal of the Transportation Research Board*, 2017.
16. Calabrese, F., M. Colonna, P. Lovisolo, D. Parata and C. Ratti. Real-time Urban Monitoring using Cell Phones: A Case Study in Rome. *IEEE Transactions on Intelligent Transportation Systems*. Vol. 12. No. 1, 2011, pp. 141-151.
17. Wang, P., T. Hunter, A. M. Bayen, K. Schechtner and M. C. Gonzalez. Understanding Road Usage Patterns in Urban Areas. *Scientific Reports*. Vol. 2, No. 1001, 2012.

18. Wang, J., D. Wei, K. He, H. Gong and P. Wang. Encapsulating Urban Traffic Rhythms into Road Networks. *Scientific Reports*. Vol. 4, No. 4141, 2014.
19. Alexander, L., S. Jiang, M. Murga and M. Gonzalez. Origin-Destination Trips by Purpose and Time of Day Inferred from Mobile Phone Data. *Transportation Research Part C: Emerging Technologies*, Vol. 58, 2015, pp. 240-250.
20. Ma, J., F. Yuan, C. Joshi, H. Li and T. Bauer. A New Framework for Development of Time-Varying O-D Matrices based on Cellular Phone Data. Presented at the 4th TRB Innovations in Travel Modeling Conference, Tampa, FL, May 2012.
21. Iqbal, M. S., C. F. Choudhury, P. Wang, M. Gonzalez. Development of Origin-Destination Matrices using Mobile Phone Call Data. *Transportation Research Part C: Emerging Technologies*, Vol. 40, 2014, pp. 63-74.
22. Toole, J. L., S. Colak, B. Sturt, L. P. Alexander, A. Evsukoff, M.C. Gonzalez. The Path Most Traveled: Travel Demand Estimation using Big Data Resources. *Transportation Research Part C: Emerging Technologies*, Vol. 58, 2015, pp. 162-177.
23. Marzano, V., Papola, A., Simonelli, F. Limits and Perspectives of Effective O-D Matrix Correction using Traffic Counts. *Transportation Research Part C: Emerging Technologies*, Vol. 17, 2009, pp. 120-132.
24. Daly, A., J. Fox and J. Tuinenga. Pivot-Point Procedures in Practical Travel Demand Modeling. Presented at the 45th Congress of the European Regional Science Association, Amsterdam, The Netherlands, August 2005.
25. Fox, J., A. Daly and B. Patruni. Enhancement of the Pivot Point Process used in the Sydney Strategic Model. Bureau of Transport Statistics, Transport for New South Wales, 2012.

NOTICE

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The United State Government assumes no liability for its contents or use thereof.

The United States Government does not endorse manufacturers or products. Trade names appear in the document only because they are essential to the content of the report.

The opinions expressed in this report belong to the authors and do not constitute an endorsement or recommendation by FHWA.

This report is being distributed through the Travel Model Improvement Program (TMIP).

U.S. Department of Transportation
Federal Highway Administration
Office of Planning, Environment, and Realty
1200 New Jersey Avenue, SE
Washington, DC 20590

October 2017

FHWA-HEP-16-078

U.S. Department of Transportation
Federal Highway Administration