Understanding Traditional Origin-Destination Data: A Survey

OCTOBER 2017

U.S. Department of Transportation
Federal Highway Administration

TMIP
Better Methods. Better Outcomes.

<table>
<tr><td>1. Report No.
FHWA-HEP-19-046</td><td>2. Government Accession No.</td><td colspan="2">3. Recipient’s Catalog No.</td></tr>
<tr><td colspan="2" rowspan="2">4. Title and Subtitle
Understanding Traditional Origin-Destination Data: A Survey</td><td colspan="2">5. Report Date
October 2017</td></tr>
<tr><td colspan="2">6. Performing Organization Code</td></tr>
<tr><td colspan="2">7. Authors
Vince Bernardin, Mark Bradley, Robert Chamberlin, Clint Daniels, Mark Fowler, Joel Freedman, Maren Outwater, Kaveh Shabani, Colin Smith, William Woodford,</td><td colspan="2">8. Performing Organization Report No.</td></tr>
<tr><td colspan="2" rowspan="2">9. Performing Organization Name and Address
RSG
55 Railroad Row
White River Junction, VT 05001</td><td colspan="2">10. Work Unit No. (TRAIS)</td></tr>
<tr><td colspan="2">11. Contract or Grant No.
DTFH61-12-D-00013</td></tr>
<tr><td colspan="2" rowspan="2">12. Sponsoring Agency Name and Address
United States Department of Transportation
Federal Highway Administration
1200 New Jersey Ave. SE
Washington, DC 20590</td><td colspan="2">13. Type of Report and Period Covered
June 2017-October 2017</td></tr>
<tr><td colspan="2">14. Sponsoring Agency Code
HEPP-30</td></tr>
<tr><td colspan="4">15. Supplementary Notes
The project was managed by Task Manager for Federal Highway Administration, Sarah Sun, who provided detailed technical directions.</td></tr>
<tr><td colspan="4">16. Abstract
Many transportation planning models rely on estimates of origin-destination (OD) travel flows for inputs or validation. Conventional 4-step travel demand models estimate origins and destinations using trip generation and trip distribution approaches. Given the importance of the OD matrix in this type of modeling, several methods and datasets have been developed for estimating OD travel flows. This volume reviews 10 traditional approaches, surveys, or datasets that can be used to estimate OD travel flows: household travel diaries; establishment survey data; external station surveys; intercept surveys; the Longitudinal Employer-Household Dynamics dataset; the Census Transportation Planning Package; visitor survey data; on-board transit survey data; Bluetooth OD estimation; and matrix estimation calibrated to traffic counts. For each approach, the challenges and shortcomings are highlighted. All the approaches have limitations including limited survey sample sizes, high cost of administration, and multiple sources of statistical bias. Analysts should understand the shortcomings in each of these traditional approaches to OD estimation and consider additional methods to address these limitations.</td></tr>
<tr><td colspan="2">17. Key Words
Origin-destination data, origin-destination estimation, travel surveys, household surveys. trip diary surveys, transit surveys, Big Data for transportation.</td><td colspan="2">18. Distribution Statement
No restrictions.</td></tr>
<tr><td>19. Security Classif. (of this report)
Unclassified</td><td>20. Security Classif. (of this page)
Unclassified</td><td>21. No. of Pages
102</td><td>22. Price
N/A</td></tr>
</table>

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)

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)

Understanding Traditional Origin-Destination Data: A Survey

Original: October 2017
Final: October 2017

Prepared for:

Federal Highway Administration

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviations

AASHTO	American Association of State Highway and Transportation Officials
AB	activity-based (model)
ABS	address-based sampling
ACS	American Community Survey
APC	automated passenger counting
BATS	Bay Area Transportation Study
BLS	Bureau of Labor Statistics
CAPI	computer-assisted on-site personal interviews
CATI	computer-assisted telephone interview
CFS	Commodity Flow Survey
CTPP	Census Transportation Planning Products
FHWA	Federal Highway Administration
HBS	home-based school
HBW	home-based work
HH	household
IPF	iterative proportional fitting
JTW	journey-to-work
LEHD	Longitudinal Employer-Household Dynamics
LODES	LEHD Origin-Destination Employment Statistics
MAC	media access control
MAG	Maricopa Association of Governments
MPO	metropolitan planning organization
MTC	Metropolitan Transportation Commission
MWR	Multiple Worksite Report
NHBO	non-home-based-other
NHTS	National Household Travel Survey
OD	origin-destination
ODME	origin-destination matrix estimation
OPM	Office of Personnel Management
PA	productions and attractions
PCF	Personal Characteristics File
PUMA	Public Use Microdata Area
QCEW	Quarterly Census of Employment and Wages
RP	revealed preference
SANDAG	San Diego Association of Governments
SP	stated preference
SSN	Social Security number
STOPS	Simplified Trips-on-Project Software
TAZ	Traffic analysis zones
TNC	transportation network company
UI	Unemployment Insurance
VIUS	Vehicle Inventory and Use Survey
VMT	vehicle miles traveled
VOT	value of time
VOTTS	value of travel time savings
WAC	workplace area characteristics
WIM	weigh-in-motion

Executive Summary

Many travel demand models rely on estimates of origin-destination (OD) travel flows for inputs or validation. Conventional 4-step travel demand models estimate origins and destinations using trip generation and trip distribution approaches. Given the importance of the OD matrix in this type of modeling, several methods and datasets can be used to support the estimation of OD travel flow.

This volume reviews 10 traditional approaches, surveys, or datasets that can be used to estimate OD travel flows:

- Household travel diaries.
- Establishment survey data.
- External station surveys.
- Intercept surveys.
- Longitudinal Employer-Household Dynamics dataset.
- Census Transportation Planning Package.
- Visitor survey data.
- On-board transit survey data.
- Stated preference OD data.
- Bluetooth OD estimation.

This volume includes a concluding chapter on using traffic counts in the OD estimation process, which also references Volume 4 in this series. Volume 4 provides a more thorough investigation of the OD Matrix Estimation (ODME) process.

A key objective of this volume is to provide guidance on how these various data sources can be used to validate OD estimates in travel demand models. To this end, for each approach, this volume provides background on how the data are collected and processed into OD estimates. Further, this volume assesses the suitability of the data—the challenges and shortcomings—for OD estimation.

Each approach has limitations, including low response rates, limited survey sample sizes, high cost of administration, and multiple sources of statistical bias. As a result, the OD estimates are burdened with aggregation biases. These biases include sparse matrices with respect to zones, coarse time periods, aggregated trip purposes, and limited purpose segmentation. Analysts should understand the issues associated with each of these traditional approaches to OD estimation and consider additional methods to address these shortcomings.

Combining data from other sources, including passive “Big Data” sources, can help address some of these shortcomings. Volumes 2, 3, and 4 from this series will address issues associated with using passive data for OD estimation. While passive data provide more temporal and spatial detail, these data also contain biases—both known and unknown. Further, passive data do not include details that are important to travel modeling like travelers’ socioeconomic characteristics.

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. This volume is not meant to be a comprehensive review of the so-called traditional methods of estimating origin-destination (OD) travel flows. This report provides a high-level synthesis of methods that have been used traditionally to estimate OD flows.

1.2 Acknowledgments

This volume is a collaboration between transportation professionals at FHWA, FTA, and RSG.

1.3 Introduction and Overview

OD data describe travel patterns in a study area of interest. Understanding the starting and ending points of trips, by mode of travel, is extremely valuable in transportation planning models and applications. OD data are typically expressed as "OD matrices" that specify the number of people, cars, transit riders, or other travel units that complete trips between one zone and another zone. OD matrices are a critical input to many travel demand models. For conventional travel demand models, OD matrices are estimated internally as the outputs of the trip distribution stage of the process.

For most study areas, precisely measuring origins and destinations is impractical and prohibitively expensive. For this reason, multiple survey methods have been developed to estimate OD matrices for different modes, times of day, and geographic scopes. Each of these methods has strengths and limitations. This volume provides an overview of traditional data sources underlying OD estimation to highlight the strengths and weaknesses of these data sources. Every approach to estimating OD data has limitations that, if utilized without knowledge of those limitations, can lead to erroneous assumptions about data quality and integrity.

This volume reviews 10 traditional approaches, surveys, or datasets that can be used to estimate OD travel flows:

- Household travel diaries.
- Establishment survey data.
- External station surveys.
- Intercept surveys.
- Longitudinal Employer-Household Dynamics dataset.
- Census Transportation Planning Package.
- Visitor survey data.
- On-board transit survey data.
- Stated preference OD data.

- Bluetooth OD estimation.

This volume includes a concluding chapter on using traffic counts in the OD estimation process, which also references Volume 4 in this series. Volume 4 provides a more thorough investigation of the OD Matrix Estimation (ODME) process.

Each section introduces the approach, describes the data collection system, discusses the analytical processes for preparing OD estimates, and assesses the strengths and shortcomings of the approach for OD estimation.

OD estimates based on survey approaches should be reviewed for accuracy and completeness. Inaccurate and incomplete survey results either need to be corrected or discarded from the final database. A smartphone-based GPS travel survey offers the greatest opportunity to capture errors in real time and prompt users for correction. Pencil and paper recall surveys are the most prone to poor recall and require the most data cleaning. The following example data checks are illustrative of what could be run to verify and clean the data.

- **Locations and Addresses Validation**: Ensure trips have origins and destinations at valid locations.
- **Purpose and Location Consistency**: Verify that the trip purpose and the destination are consistent. For example, ensure that dining trips have a destination at a restaurant.
- **Travel Time Validation**: Ensure that the travel time reported between origin and destination is reasonable based on the mode, route, and time of day.
- **Dwell Time Validation**: Verify that time spent at a location is reasonable based on the trip purpose and type of location.
- **Trip Chaining Consistency**: Verify that the reported departure time is after the arrival time of the previous trip destination. This also includes verifying that the destination of the previous trip is consistent with the origin of the current trip.
- **Time-of-Day Validation**: Ensure that trips are occurring at logical times in the survey. For example, it is unlikely for an elementary school trip type to leave home late at night.

Regarding survey data, important themes that are raised in this volume include the following:

- Are the data collected from a representative sample or from a targeted sample? If the sample is not representative, what efforts should be made to correct for sampling bias?
- For any sample, what weighting procedure should be used to expand the sample to represent the total population?
- How can Big Data sources complement traditional survey methods?

2.0 Household Travel Survey Data

2.1 Background

A household travel survey collects information about individuals (e.g., age and employment status), their household (e.g., size, structure, and income), and a diary of their trips on select days. For each trip, the survey collects the start and end locations, the departure and arrival time, the mode of travel, the number and type of fellow travelers, and the purpose and duration of the activity at the trip destination. The information collected is like a trip origin-destination (OD) survey, but a household travel survey is more comprehensive and captures information for *all* trips made by all household members during the study period.

As with other OD estimation methods, data obtained through household travel surveys can be used in travel demand models to estimate and calibrate distribution or destination choice models, as well as for validating trip matrices estimated through other methods. As with all traditional OD estimation methods, passive data sources provide new datasets that can be combined with traditional methods for validation.

2.2 Introduction

In the United States, the National Household Travel Survey (NHTS) has been conducted in 1969, 1977, 1983, 1990, 1995, 2001, 2009, and 2016.[1] The NHTS monitors long-term national travel trends; however, because the survey covers the entire nation, it typically does not have a large enough sample for any one region to support regional travel demand modeling. In part to address this limitation, some states and regions have purchased their own “add-on” samples in recent NHTS surveys to support statewide or regional modeling.

The 1956 Chicago Area Transportation Study home interview survey was one of the first regional household travel surveys,[2] which is considered pioneering work in this field. In 1965, one of the first regional household travel surveys was conducted in the San Francisco Bay Area by the Bay Area Transportation Study (BATS) Commission, which later became the Metropolitan Transportation Commission. The BATS travel survey has continued, supporting a long tradition of disaggregate travel demand modeling in the Bay Area. Since then, most large metropolitan planning organizations (MPOs) have undertaken similar studies, conducting regional or statewide household travel surveys approximately once per decade. Until 2000, most regions conducted their surveys concurrently with the decennial census. This alignment supplied necessary sociodemographic data for survey weighting. Now, with the annual American Community Survey (ACS) replacing the decennial census long form, it is less important for surveys to be timed around the census. Instead, the timing of surveys depends more on an MPO’s (or department of transportation’s) region-specific needs around travel models and funding.

In larger regions, the sample size for a regional household travel survey is often between 0.4% to 0.6% of the region’s households (e.g., 4,000 to 6,000 households for a region with 1 million households). Smaller regions are more likely to have modestly higher sample sizes as a

[1] National Household Travel Survey: Understanding How People Get from Place to Place

[2] The Chicago Area Transportation Study: Creating the First Plan (1955-1962)

percentage of the region's households. Participating households often report all travel for an assigned period, typically one weekday—usually in the spring or the fall to avoid the "atypical" summer and winter vacation/holiday periods.

By way of example, a survey using traditional methods that obtains 5,000 households, with 2.5 people per household and 4 trips per person-day, will result in around 50,000 individual trip records. A typical MPO might have 2,000 traffic analysis zones (TAZs) in its travel demand model, producing 2,000 x 2,000, or 4 million possible origin-destination (OD) pairs. So, even if every survey trip was between a new OD pair, the survey would only cover 50,000/4,000,000, or 1.25% of all possible OD pairs.

In reality, many MPOs include more than 2,000 TAZs and many survey trips are between the same OD pairs. As a result, the OD data from a household travel survey may often cover well below 1% of all possible OD pairs in a region. Rather than relying on the household travel survey to provide comprehensive OD data directly, the data are used to estimate and calibrate travel demand models that are then used to synthesize OD trip matrices by travel mode and time of day. The form of the travel demand models has evolved in some regions over the last two decades—from trip-based to tour-based to activity-based (AB). However, the basic process of using zone-based travel demand models (integrated with static, equilibrium network traffic assignment) to synthesize OD trip matrices has remained essentially the same. In all methods, even including the most advanced AB models, the limited OD coverage of the underlying household travel survey data is a constraint that limits travel demand modelers' ability to accurately model and predict spatial trip OD patterns.

The following sections describe trends in household travel surveys, emphasizing new methods that increase the volume or accuracy of collected trip OD data.

2.3 Overview of Household Travel Survey Methods

The main topics covered in this section include the following:

- Data contents and accuracy.
- Data completeness and nonresponse.
- Combining household travel data collection methods.
- Sampling and weighting issues.
- Improving OD coverage and accuracy.

2.3.1 Data Contents and Accuracy

A household travel survey typically collects the following data:

- Household characteristics.
- Person characteristics of each household member.
- Vehicle characteristics of each household vehicle.
- Trip characteristics for every trip made by each person.

- Modes and routes used for each trip segment.

For use in AB model estimation and calibration, further survey data processing groups trips into tours (home-based and work-based trip chains) and groups tours into person-days and household-days. Data accuracy has grown in importance given the more advanced model systems now used in most regions:

2.3.1.1 *Locations*

In addition to TAZs, many AB models use micro analysis zones or parcels for a more detailed description of land use and more accurate distances for short trips. Accurate geocoding of trip ends is important. At a minimum, and to enable higher accuracy, the retrieval of travel data should use a real-time map-based interface, such as Google Maps, or smartphone-based GPS data collection.

2.3.1.2 *Times*

Capturing accurate trip departure and arrival times (which are the same as activity start and end times) is also important. These times help estimate or calibrate activity-scheduling models and help impute travel times for data quality control. Rounding or poor recall produce inaccuracies in self-reported travel times from trip diaries or from memory. Smartphone-based GPS data collection can more accurately capture travel times.

2.3.1.3 *Modes and Routes*

For travel modeling, it is important to capture all modes used for a given trip, including access and egress modes, transfers, and vehicle/service types for all legs of transit trips. For auto trips, it can be useful to capture parking details, and parking locations for any trips where the parking place is a nonnegligible walking distance from the destination. Smartphone-based GPS travel surveys require some type of prompted recall to capture self-reported mode use. (In some cases, it is possible to impute the mode used from the GPS trace data, particularly for walking as an access or egress portion of a trip.) Even when self-reported mode data are captured, GPS traces can be used to impute or verify locations and other data in trips that include multiple modes (e.g., transit boarding and alighting points). Smartphone-based GPS data also can be used to observe the route taken for any trip, including auto trips, although detailed analysis methods are necessary to impute entire routes in cases where the trace data are ambiguous.

2.3.1.4 *Activities*

Travel demand models that segment activities across types require accurate data on activity purposes. Traditional survey methods typically predefine the usual work or school locations for all household members, and this information can be used to verify or impute the purpose for most work and school activities. For smartphone-based GPS travel surveys, prompted recall helps identify respondents' self-reported activity purposes for all trips; imputing activity purposes post hoc based on land-use data is often less accurate. Lists of activity types shown to respondents should be comprehensive enough to capture all activity types that may be of interest. For example, if the data will be used to estimate physical activity, then separating exercise/active recreation from other types of recreation is useful.

Some surveys have asked respondents for details on their in-home activities, including multiple activities for each at-home time span. A primary focus of in-home activities is to identify substitutions for travel such as online shopping, banking, working, and classes. However, details of in-home activities are not needed for any models currently in use. These questions also increase respondent burden, may provoke feelings of intrusiveness, and deliver inconsistent results (some people provide more detail than others). Broad questions about in-home activities at the full-day level can achieve similar results, such as "Did you spend any time doing paid work while at home? If so, how much time did you work at home?"

2.3.1.5 *Household Joint Travel*

Some AB model systems explicitly predict joint travel and activities across household members. In such cases, travel data for different household members who travel together should be accurate and consistent. Typically, people are also asked to identify other household members who traveled on each trip, and these data can be cross-referenced for consistency. Ideally, online travel diary data retrieval software contains a "trip copy" feature that automatically copies trip details to all other travelers in the household to ensure data consistency. For children and some adults, some trips are provided by proxy, and a "copy trip" feature helps "jump start" the process of proxy reporting, as any trips that were made together with the respondent are already included for the other household member(s).

2.3.1.6 *Traditional Travel Diary and Smartphone-Based GPS Survey Methods: Pros and Cons*

Travel surveys have used GPS data collection for the last decade or more—typically for a subsample of respondents to compare and adjust traditional travel survey data. Until recently, GPS data collection in MPO travel surveys relied on purpose-specific "black box" GPS tracking devices mailed to respondents. Respondents carried the devices for a specified period and then returned them. The GPS data collection was typically done for a small subsample (e.g., 10%) to calculate trip rate bias correction factors. There have also been a handful of "GPS-only" surveys, including those in Cincinnati and Cleveland, as well as a large "GPS-only" sample in the San Francisco Bay Area as part of the 2012 California Household Travel Survey.

Drawbacks to using standalone GPS devices include the following:

- Buying and shipping the devices is expensive (particularly since some fraction of devices are never returned).
- The devices only capture time and location. Survey questions about modes used, activity purposes, and fellow travelers require transferring the GPS trace data online and then having the respondents do prompted recall. Expense, burden, and recall error are top concerns, especially if the elapsed time between GPS capture and the online recall survey is long.
- Devices are often forgotten or left at home. People are not used to carrying standalone GPS devices and often leave the devices at home or in their vehicles. The devices must also be charged, which not all respondents remember to do.

Smartphone-based GPS travel survey methods address many of the issues with standalone GPS devices. These applications ("apps"), installed on respondents' own smartphones, capture travel for multiple days—seven days is a typical data collection period. Respondents report all details of their trips via the app when it is convenient. Respondents who own smartphones are used to carrying the devices, keeping them charged, and frequently checking for new notifications. For these reasons, smartphone-based GPS travel survey data collection is quickly becoming the standard method for household travel surveys.

Table 1 summarizes recent household travel surveys in the United States, and most of the examples from the last two years employ smartphone-based GPS travel survey data collection. The market for these smartphone apps is maturing rapidly. These apps include new data prompt features and data processing tools to maximize data quality and completeness. Costs are also decreasing—many smartphone-based GPS travel surveys are now comparably priced with traditional travel diary data collection methods. In fact, with the ability to collect multiple travel days, smartphone-based GPS travel survey data collection is already *less* expensive than traditional travel survey data collection on a per-travel-day basis.

Table 1. Summary of recent regional household travel survey main collection methods.

Region (Client)	Year	HH In Region	HH In Sample	% of HH In Sample	Main Method
Minneapolis (Met Council)	Upcoming 2018	1,500,000	7,500	0.50%	Smartphone-based (traditional travel survey for nonsmartphone owners)
Chicago (CMAP)	Upcoming 2018	3,000,000	12,000	0.40%	Smartphone-based (traditional travel survey for nonsmartphone owners)
Washington, DC (MWCOG)	2017/18	2,500,000	15,000	0.60%	Traditional travel diary
San Diego (SANDAG)	2106/17	1,100,000	6,000	0.55%	Smartphone-based (traditional travel diary for nonsmartphone owners)
Columbus (Ohio Department of Transportation)	2016/17	500,000	3,000	0.60%	Smartphone-based (loaned phones or traditional travel diary for nonsmartphone owners)
Phoenix (MAG)	2016/17	1,500,000	7,000	0.47%	Smartphone-based (traditional travel diary for nonsmartphone owners)

Region (Client)	Year	HH In Region	HH In Sample	% of HH In Sample	Main Method
Research Triangle (ITRE)	2016	650,000	4,200	0.65%	Traditional travel diary; smartphone subsample
Seattle (PSRC)	2014-15; 2017	1,500,000	6,300	0.42%	Traditional travel diary; smartphone subsample

In most smartphone-based GPS travel surveys, adults who do not own smartphones provide their travel via traditional travel diary methods (online or by CATI). The state of Ohio has elected for their project to provide smartphones to low-income households where some or all adults do not have a smartphone but are willing to participate using one. The primary drawbacks to loaning smartphones mirror the drawbacks encountered in previous studies that used standalone GPS devices. For example, drawbacks to loaning smartphones to participants include the expense of purchase, delivery, and retrieval. These costs mean fewer households can be surveyed overall for a given project budget. Also, loaned phones are easier to forget or misplace since many participants requiring loaners also carry a simpler cellphone or do not carry a cellphone at all (and may be unfamiliar with the technology). As a result, some trips are missed and the survey records more “stay-at-home” days.

Compared to “standard” travel data collection methods using travel diaries and subsequent telephone or online data retrieval software, the advantages of using smartphone-based GPS travel surveys include the following:

- Reduced respondent burden as respondents do not have to remember trip times, enter addresses, or locate trip ends on maps. Instead, travel is automatically captured and displayed for them on a map within the smartphone app.
- Improved accuracy of collection of trip-end locations and times. (Fellow travelers will register the same locations and times.)
- Improved trace data to impute routes used, changes of mode, access and egress times, waiting times, and parking locations, among others. These data can be visualized in subsequent analyses.
- Enhanced ability to capture multiple days of travel with little additional cost. Due to lower respondent burden, many respondents who complete one day go on to complete all the days requested by the study (e.g., 3 or 7 days).
- Decreased cost per respondent (eventually). Smartphone-based GPS travel survey data collection already costs less per collected travel day (the ability to collect up to seven or more travel days). Currently, the costs for smartphone-based GPS travel survey methods are comparable to those of traditional travel diary methods on a per-respondent basis; these costs are less on a per respondent-day basis when capturing multiple days per person. (With either approach, the major survey cost is in recruiting the respondents, not in collecting the trip data.)

- Diminished age differences in responses. Young respondents (millennials) seem to be just as willing as other age groups to participate; this group is often underrepresented in traditional travel diary surveys.

Disadvantages of smartphone-based GPS travel surveys when compared to traditional travel diary survey methods include the following (most of which may diminish over time):

- Lack of universal ownership. Approximately 77% of US adults own smartphones, which means not everyone has one. And children with smartphones may not be allowed by their parents to use them for the survey. Obtaining a fully representative sample still requires loaning smartphones to respondents or employing a mixed approach with both smartphone-based GPS travel survey methods and traditional travel diary methods.
- Lack of uniform smartphone devices and features. This manifests in several ways. For example, although the Apple iPhone has limited models from one manufacturer, there are hundreds of Android phones across a variety of major companies such as Samsung, Motorola, and HTC. The accuracy and quality of the specific sensors and battery life within these phones vary. In general, the newer the phone model, the better the battery life and the accuracy of the passive data. Similarly, each specific app that collects this travel survey data is expected to have different proprietary technology for how to preserve battery power during data collection.
- Data privacy concerns. Respondents who are concerned about data privacy may be more concerned with automatic GPS tracking (although it seems likely that such respondents would also avoid completing travel diaries).
- Additional time required for software best practices and approvals. Each version of the app must first be registered with the App Store (iOS) and Google Play (Android). Updates to the app are published in the stores and in the case of Apple sometimes require additional time for approval. This precludes last-minute changes to app functionality.

2.3.2 Data Completeness and Nonresponse

2.3.2.1 *Item Nonresponse*

Item nonresponse refers to missing data for specific questions in cases where the survey data are otherwise complete. Missing data can be problematic for data weighting and expansion, and this makes the data more difficult to use in model estimation and calibration. On the other hand, some respondents find questions about income and race/ethnicity to be intrusive. While item nonresponse is difficult to avoid in mailed-back self-completion surveys, it is easier to control in online, smartphone, or other computer-based surveys that can be validated in real time.

One approach allows item nonresponse—but only for income and race/ethnicity. Complete data are typically required for all other data items. In practice, between 5% to 15% of respondents typically refuse to provide their household income. The percentage can be reduced somewhat by following the detailed income question with another question that asks about income but does so within fewer, broader categories (e.g., 5 answer choices instead of 10 detailed answer choices) for those who refuse to answer the more detailed question.

OD data collection prioritizes complete information and geocoding for all trip-end addresses and locations. Previous household travel surveys encountered issues with missing or inaccurate trip-end location data. Missing location data is rarer now with online, computer-assisted telephone interview (CATI), and smartphone-based GPS data collection. (That said, accuracy appears even better with smartphone-based GPS travel survey methods than with online or CATI methods.) Paper-based mail-back surveys are the most deficient with respect to missing location data and other key data items. The deficiencies with paper-based mail-back survey methods have contributed to their precipitous decline in use in the United States.

2.3.2.2 *Missing Trips*

Nonreported (or missing) trips are one of the more problematic issues affecting traditional travel diary surveys. Approximately 10% to 20% of respondents report not making any trips on their assigned travel day when completing traditional travel diary surveys. While many of those cases are accurate, there is also compelling evidence that many of them are “soft refusals.” Soft refusals describe participants who participate by identifying the fastest path to completing the survey (by not recording their travel). Recent evidence from smartphone-based GPS travel surveys suggests that the true frequency of people remaining at home all day without traveling is 30% to 50% lower than that reported in traditional travel diary surveys. Consistent with this finding, AB models that are calibrated to unadjusted traditional travel diary data typically produce too few trips when compared to external validation data; adjusting the fraction of “stay-at-home days” is one of the most efficient ways of calibrating the models. To address this issue, GPS data have historically been collected for a subsample of respondents and used to adjust the trip rates for calibration. (As discussed in the weighting section below, newer data collection methods allow for more accurate forms of adjustment in calibration and estimation.)

In addition to the “stay-at-home” as “soft refusal” issue, respondents using traditional travel diary survey methods often underreport certain types of trips when compared to smartphone-based GPS travel survey methods. The trips that are underreported tend to be shorter trips—both walk trips and car trips. Trip-diary respondents sometimes fail to report intermediate stops on multistop auto tours, producing too few trips and a bias toward longer trips and more-distant OD pairs.

2.3.2.3 *Missing Household Members and Proxy Reporting*

Some household travel surveys (including the NHTS) have tried not requiring that every household member provide complete travel data for the assigned travel day. In these instances, some percentage of households are called “complete” even with missing household travel data. However, most regional or statewide household travel surveys require complete travel data for the assigned day for either all household members or all household members age five or older. If an AB model system contains models of joint travel across household members, allowing incomplete households would bias the data and any models estimated or calibrated using the data. In cases where children travel together with other household members, their travel is captured indirectly via other persons’ trips. Since young children less frequently travel on their own, missing data are limited to instances when a young child is accompanied by a nonhousehold member, such as a nanny, a grandparent, or the parent of another child. For older children—particularly young teenagers—the number of trips they make on their own is more substantial. A typical proxy reporting approach is to also ask the adults to report any trips that their children make without them (since joint-trips will have been captured in the adults' diaries). While proxy reporting is likely to capture the more regular trips and home-based trips, such as trips to/from school or friends' houses, it may miss shorter intermediate stops that children make during those tours. Missing trips are also an issue for traditional methods when proxy reporting is allowed for adults, where one adult reports travel for another household adult with that adult not present. (This does not occur with smartphone survey methods given that each person uses his or her own phone). Proxy reporting requires flagging any travel days reported by proxy in the data to permit bias corrections.

2.3.3 Combining Household Travel Data Collection Methods

This section discusses methods for combining traditional travel data with smartphone-based GPS travel survey data to develop useful OD estimates.

Two approaches can help combine smartphone-based GPS travel survey methods and traditional travel diary survey methods. One approach has a small subset of respondents use smartphones while most of the other respondents use traditional travel diaries. This approach is analogous to past subsample approaches using custom GPS devices. However, in this case, smartphone-based GPS travel survey respondents do not need to also complete (redundant) travel diaries for the same days. The second approach has smartphone owners use the app while people without smartphones use traditional travel diaries; this includes children under a certain age who do not have their parents' permission to use their smartphone for the survey. This second approach requires the following considerations:

- Should smartphones be loaned or given to some or all adults who do not own them?
- Should people who own smartphones be given the option to use the traditional diary method instead if they prefer?
- Should all adults in the household be required to use the same method (smartphone or diary), or should the survey allow mixed methods across household adults within a given household?

Allowing multiple methods across members of the same household is a complex issue because smartphone-based GPS travel survey data collection permits multiple days of data collection. Traditional travel diary data collection is rarely done for more than one or two days, which reduces the complexity of multiple methods. For an AB model that simulates joint travel and activities, some models can only be estimated or calibrated using data for days when all household members' (or at least all household members old enough to travel alone) travel data is complete.

2.3.4 Sampling and Weighting Issues

2.3.4.1 *Address-Based Sampling*

Regardless of whether traditional travel diary methods or smartphone-based GPS travel survey methods are used to retrieve household travel survey data, the costliest part of the survey process is recruitment. In the past couple of decades, the most common household travel survey recruitment method has been random-digit dialing. Over the last decade or more, with more people screening their incoming phone calls and foregoing landline phones, the commonly accepted practice has transitioned to address-based sampling (ABS), which typically relies on the following steps:

- Specify the sampling rates by geographic areas such as ZIP Codes, census tracts, or census block groups.
- Purchase a random sample of addresses with the specified geographic distribution from the US Postal Service's address database.
- Mail a prenotice introducing the survey to the purchased addresses. This increases the likelihood that households will notice a second mailing with the detailed survey materials. (Prenotice postcards are often mailed in interval batches to monitor response rates and facilitate sample adjustment.)
- Mail a second invitation packet with more detailed information about the survey.
- Mail reminder postcards to those who have not responded. Phone calls (if available and matched to addresses) can also help recruit respondents.
- Assign a (starting) travel day if a respondent completes the recruitment survey (via a URL to an online survey or by phone) and agrees to participate. (For a smartphone-based GPS travel survey, respondents are sent a text or e-mail with a URL and password to download and register the app.)

Although ABS can potentially reach any household with a mailing address, and there is no better alternative currently available, it still has some limitations:

- The sample excludes group quarters, such as college dormitories or military bases.
- Younger people and people with lower incomes tend to change addresses more often, increasing the likelihood the address is out of date.
- Response rates are often low, with between 3% and 10% representing a typical range. (Mailing fewer initial invitations—but more follow-up reminders—increases response rates but costs more; incentives also affect response rate as discussed below.)
- Low response rates often signal a high potential for some degree of selective nonresponse bias. Older households and higher-income households respond more frequently, and younger households and lower-income households respond less frequently. (Smartphone-based retrieval for the subsequent travel portion helps address the age bias but not the income bias.)

2.3.4.2 *Targeted and Compensatory Oversampling*

Household travel surveys have historically prioritized obtaining a representative and unbiased sample. A corresponding emphasis has also been placed on so-called "probability sampling," whereby the relative probability of including any given household in the sample is assumed to be known based on population data. These probabilities can then be used to expand the sample to represent the full population.

From a modeling standpoint, targeting rarer demographics and travel behavior groups, and capturing more of these groups in the sample, is more important than obtaining a representative sample, which—given the known biases—is unlikely. Recent surveys have recommended substantial “targeted oversampling” of groups exhibiting different travel behavior or comprising a low percentage of the general population. Examples of these groups include zero-vehicle households, low-income households, young single-person households, transit users, and bicycle users. This targeting is effective when using much higher sampling rates in census block groups that contain the highest proportions of such households or persons, according to ACS data. (Block group-level ACS data are only available for the combined 5-year ACS data tables.)

For data for AB model estimation, geographic oversampling has two purposes. First, geographic oversampling overcomes nonresponse biases anticipated using past surveys. For example, if one expects the response rates from block groups with a high percentage of low-income households to be only two-thirds of the response rate anticipated from other block groups, then 50% more invitations should be sent out to those block groups to compensate. This is meant only to provide a representative sample, so it is not “oversampling” in the traditional sense. The term “compensatory oversampling” describes this type of sampling rate adjustment.

In addition to “compensatory oversampling,” “targeted oversampling” can be used for geographic areas that are highest in terms of the types of demographic groups or mode use that is desired. For example, it is often observed in the ACS commute data that most of those who commute by bicycle live within a small minority of block groups. So, to effectively obtain more bicycle commuters in the survey sample, those block groups with the highest bicycle mode share should have an invitation rate that is higher than the “normal” invitation rate for block groups that are not a target for oversampling.

2.3.4.3 *Use of Incentives*

Household travel survey response rates are declining, and survey completion incentives can cost-effectively increase response rates. Incentives can also increase sample representativeness and are standard practice now. The NHTS offers a $20 gift card incentive and various “pre-incentives” for opening the invitation envelope and completing the initial recruitment survey. (Lower-income households are typically least likely to respond to travel surveys, and these households may be influenced most by incentives.) Incentives can be offered as cash or as gift cards to popular retailers (allowing the respondent to choose).

Incentives can be offered per household or per person, although in either case, a complete household travel day should be required to receive the incentive. Offering per-person incentives is more expensive, but this can help recruit larger households, which are also typically underrepresented in household travel survey samples. In states or regions that do not allow direct incentives, it may be possible to offer a raffle/sweepstakes type of prize, although experience has shown that these are less cost-effective than direct incentives.

2.3.4.4 *Possible Added “Convenience” Samples*

Other methods (in addition to geographic oversampling) exist for sampling groups that are underrepresented through ABS, or that have small incidences in the general population and a larger desired sample size for modeling. Convenience sampling is a nonprobability sampling technique where respondents are selected because of their convenient accessibility and proximity. A common type of “convenience sampling” is for university students. University students who live in dormitories are not usually reached with ABS, and other university students living off campus (but not with parents) typically also have low response rates. College administrators can also provide e-mail lists of students and distribute invitations for an online survey. Since recruitment is the most expensive part of a household travel survey, this low-cost recruitment can make university surveys much less expensive per respondent than household surveys. University surveys can also be made simpler and less expensive by offering only an online or smartphone response option and by only surveying students rather than households (except for students who still live at home with their families). Military base residents (and employees) can be similarly surveyed via an e-mail list provided by base administrators. An additional issue for military bases is security, which may not allow respondents to use a GPS location tracking device or otherwise report where they go while they are traveling on the base. Convenience sampling also works with vanpool, car-share, and other mode-specific membership lists.

The other main type of convenience sampling is intercept sampling for certain types of behavior of interest. Examples include the following:

- Travelers intercepted at park-and-ride lots (or via license plate photos).
- Travelers intercepted at downtown parking garages (or via license plate photos).
- Toll road/managed lane users identified via license plate photos or transponders.
- Transit users intercepted at stations or in vehicles (e.g., during an on-board survey).
- Bicyclists intercepted en route (e.g., during a bicycle count survey or OD survey).
- Nonresident visitors intercepted at hotels, convention centers, and airports, among others.

In most cases, intercept surveys (see Chapter 5.0) are done to generate their own type of data, such as count data or occupancy data for model validation. These surveys can also be used to invite additional respondents (and their households) to participate in the household travel survey to collect behavior data that are especially useful for estimating the AB model components. Intercepting a traveler while making a certain type of trip does not guarantee that they will repeat that same type of choice while participating in the household survey, but it is more likely—particularly in a multiday survey context.

Targeted, mode-specific oversampling is often necessary to obtain representative OD data for each mode of travel. This can be done through geographic oversampling or intercept sampling. Another option is to field separate OD intercept surveys for the specific mode—such as a transit on-board survey—and use that data to help weight the data and calibrate resulting models.

2.3.4.5 *Weighting the Data for Descriptive Analysis and Model Calibration*

Weighting the survey data to represent the broader population is important for descriptive analyses and for deriving model calibration targets. For model calibration, some additional adjustment of weights may be useful, as discussed in this section.

In general, weighting of household survey data follows two main steps:

1. Initial weighting based on sampling probabilities: If the sampling rates are determined geographically, then this step requires identifying each sampling area within which all households had an equal probability of being sampled; an initial weight for each area is then estimated, which equals the number of households living within the area (e.g., from the most recent ACS data) divided by the number of households in the survey sample.
2. Reweighting to match ACS-based marginal distributions: This step involves starting with the initial weight for each household and using a method like iterative proportional fitting (IPF) to simultaneously match several marginal distributions based on the most recent ACS data. (If the marginals are for large areas—counties or larger—the 1-year ACS data may be sufficient for the marginal targets.) A region with an AB model system that simulates individual persons and uses person type as a key variable should use household- and person-level marginal distributions:
    - Household residence Public Use Microdata Area (PUMA).
    - Household size.
    - Household number of workers.
    - Household income group.
    - Household vehicle ownership.
    - Household presence/absence of children.
    - Person age group.
    - Person worker status (e.g., full time, part time, nonemployed).
    - Person university student status.
    - Person ethnicity/race (more important in some regions than others).

If "missing" responses are allowed for household income or person race/ethnicity, then it is necessary to impute the data for those cases before weighting the data.

For most MPO regions, the population is large enough to accommodate separate weighting for different subregions. One approach to weighting that seems to work well and provide reasonable geographic accuracy is weighting separately within each census PUMA. PUMAs each have a similar population, helping to avoid issues with small cell sizes in weighting. When needed, adjacent PUMAs with similar demographic profiles can be combined in weighting to avoid small-cell-size issues and prevent the resulting weights from becoming too large or small in the IPF process. PUMAs have another attractive feature—they are the geographic unit available in the ACS PUMS microdata. Using the microdata instead of the published ACS tables to derive weighting targets allows more flexible target definition (one is not dependent on which tables happen to be available from the U.S. Census Bureau). Microdata use also helps address inconsistencies between ACS household-level weights and ACS person-level weights.

If the data combine smartphone-based GPS travel survey data and traditional travel diary data, then weights can be further adjusted based on a comparison of trip rates from the data types.

This approach has been applied for recent surveys with GPS subsamples, but having fully compatible data from both methods will allow weighting to be done in a more sophisticated way.

Method-specific biases in the data are often most effectively adjusted in model estimation rather than in weighting and calibration. Because the smartphone-based GPS travel survey data are fully compatible with the traditional travel diary data, the two types of data can be used simultaneously in model estimation. (That was not the case for most GPS data collected in past surveys.) A dummy variable for the traditional travel diary observations can be used in estimating any models to capture biases specific to the data collection method. (This is analogous to using a dummy variable for data reported by proxy to capture biases due to indirect reporting.)

2.3.5 Improving Origin-Destination Coverage and Accuracy

Smartphone-based GPS travel surveys have several advantages compared to traditional travel diary methods. Specifically, these newer survey methods often increase the coverage and accuracy of OD information.

2.3.5.1 *More Accurate Recording of Trip-End Locations and Times*

Smartphone-based GPS travel surveys record all times and locations, and these are generally more accurate than what respondents report. (This is particularly true for times of day and travel times, which respondents tend to round to the nearest 5, 10, or 15 minutes.) For locations, software for traditional travel diary methods has improved using the familiar Google Maps interface, and most smartphone-based apps use this same interface for respondents to help them identify their trips and provide additional information. Smartphone-based GPS travel survey data also include extra information with time and location traces en route during each trip, allowing additional analysis of route choices, imputation of mode changes, imputation of transit boarding locations, walk access times and wait times, and imputation of downtown parking locations and walk egress times.

2.3.5.2 *Multiple Travel Days per Household*

One of the key advantages of smartphone-based GPS travel data collection is that—with the reduced burden of reporting trips—most respondents are willing to provide up to 5 or 7 complete days of travel data. Recent experience with these surveys demonstrates that households who have provided at least 1 complete travel day for all members have gone on to provide an average of over 4 complete days, with over 50% completing 7 days. Given that the main cost of collecting household travel survey data is in recruitment, and that the marginal cost of each additional day of smartphone-based GPS travel data is low, the additional days of data can significantly raise the effective sample size for relatively little additional cost. One question that is often raised is the relative value of collecting multiple survey days from the same households, as compared to getting more households to each do a one-day survey. To answer this, one must first determine what percentage of trips collected on each day are unique compared to the same person's trips on previous days. In other words, how much new information does each survey day produce?

To illustrate, consider the San Diego Association of Government's recent household travel survey. There, a smartphone-based GPS travel survey collected trips for up to 7 days from approximately 4,000 households. (Additional sample participated via a traditional one-day travel survey and are not included here because they only participated for one day.) For the smartphone-based GPS travel survey data sample, a trip qualified as "unique" if the respondent made no other trips on previous survey days between the same two census blocks by the same mode at approximately the same time of day (within a 2-hour gap). Figure 1 shows that approximately 60% of home-based work (HBW) and home-based school (HBS) trips were unique on day two, but by the fifth weekday (Monday), approximately 30% of the HBW and HBS trips are still unique and had not been observed on previous days of data collection.

For other trip purposes (home-based other [HBO], non-home-based work [NHBW], and non-home-based other [NHBO]), and for total trips, between 70% and 90% of trips are still unique by the fifth weekday. (The percentage of unique trips is even higher on weekends, but most agencies do not use weekend days for modeling, so weekend trips are analyzed after weekdays.)

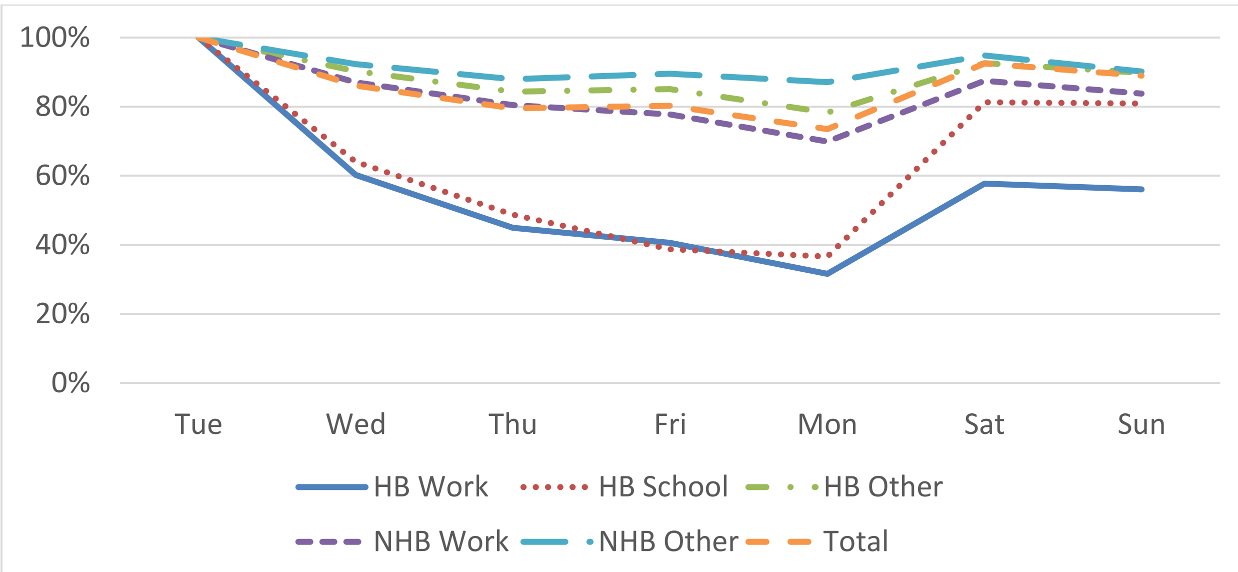


Figure 1. Percentage of unique trips on each day collected in SANDAG smartphone-based GPS travel survey data.

Figure 2 is based on the same information as Figure 1, but presents "unique" trips as a multiple of day-one trips. For HBW and HBS, 5 weekdays of data provide approximately 2.5 times as many unique trips as 1 day. For the other purposes, five weekdays of data provide more than four times as many unique trips as one day. This evidence strongly supports the value of collecting multiple days of smartphone-based GPS travel survey data, particularly given each additional day of data costs a portion of the cost that would be required to recruit another household into the survey.

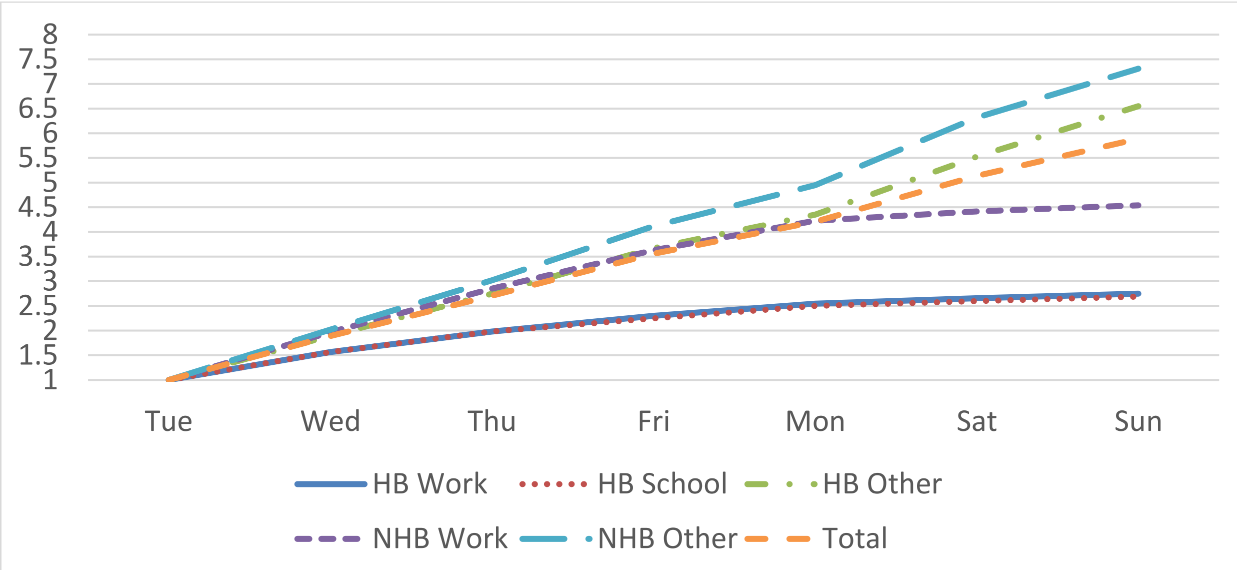


Figure 2. Cumulative number of unique trips as a multiple of Day 1 trips, based on SANDAG smartphone-based GPS travel survey data.

The household and person characteristics within a household do not change during a week. Thus, analyzing more unique trips from the same set of households does not provide as much statistical information as a larger household sample. Given that the total sample size and distribution is adequate, however, there will already be a wide spectrum of household and person types making each type of trip.

The type of analysis described for "unique trips" was also conducted for trip OD coverage. What is the gain in the number of unique OD pairs observed for each additional travel day collected? Figure 3 shows the percentage of trips on each survey day that are between OD pairs for which trips were observed on any previous trips (for the same OD pair) across the entire survey sample in previous days. For HBW and HBS trips, between 20% to 30% of trips represent new OD pairs by the fifth weekday. For HBO, NHBW, and NHBO trips, over 60% of trips represent new OD pairs. Across all trip purposes, the fraction of trips between new OD pairs is still above 50% even on the fifth weekday. Adding more days produces more spatial OD coverage. (Figure 3 shows that an even higher percentage of trips have unique ODs on weekends).

Of note is that this type of OD analysis is conducted to represent "average" travel behavior. One or two-day trip diaries do not generate a large enough sample to capture *average* behavior, given how much diversity there is from day to day. Measuring different travelers on different days does not adequately address this issue. Extending the survey period to, for example, a total of seven days is more useful for estimating average behavior.

Figure 4 shows that for HBS and HBW, 5 weekdays of data produce approximately 2.2 times as many OD pairs in the dataset when compared to surveying a single weekday. For other trip purposes, the ratio is approximately 4 times as many OD pairs after 5 days, and across all trip OD types the ratio is 3.5. The outcome is that surveying respondents for five weekdays will provide at least three times the OD coverage as a single-day travel survey.

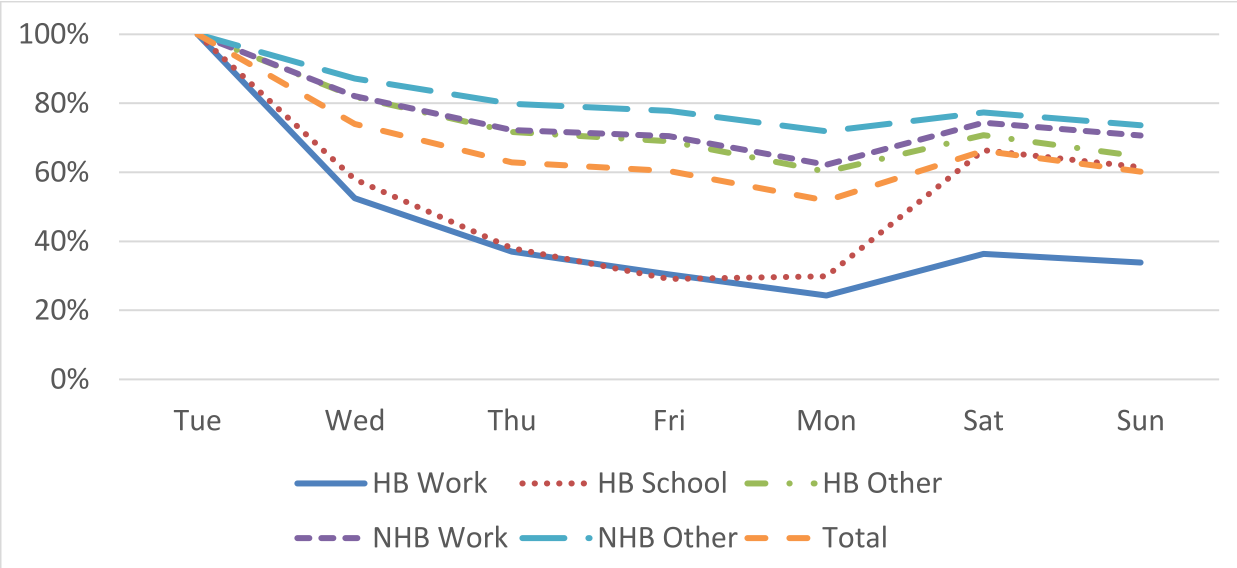


Figure 3. Percentage of trips on each day with unique OD pairs from SANDAG smartphone-based GPS travel survey.

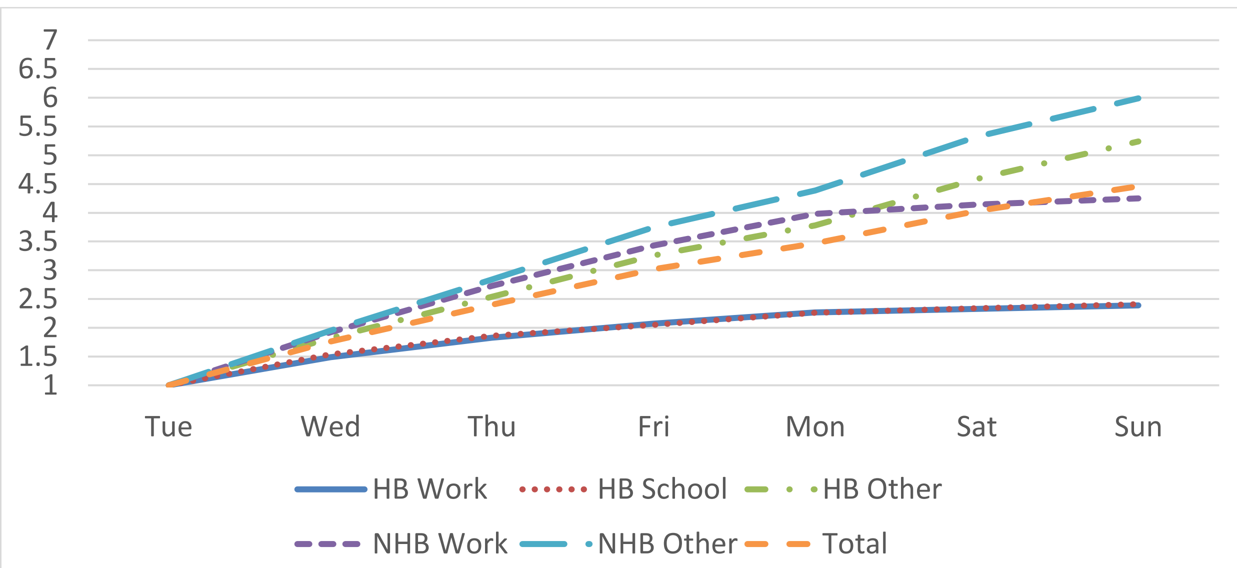


Figure 4. Cumulative number of unique OD pairs as a multiple of Day 1 OD pairs, based on SANDAG smartphone-based GPS travel survey data.

2.4 Assessment of Data Suitability

This section discusses the strengths and shortcomings of household travel survey OD information and summarizes recent trends and further opportunities to improve this data source.

2.4.1 Strengths

The fundamental strength of household travel survey data is the richness of the data it provides, including all trips made by all household members over a given survey period. This richness makes the data suitable for estimating and calibrating travel demand models, ranging from simpler trip-based models to advanced AB models.

2.4.2 Shortcomings

The most significant shortcoming of household travel surveys regarding OD data is the limited sample size that can be surveyed within a feasible budget. A typical household travel survey sample only includes approximately 0.5% of the households in a region, and the trips that are collected typically cover fewer than 1% of all possible OD zone pairs (depending on the number and size of zones used in the region). This lack of OD coverage limits the amount of detail and level of accuracy of any destination choice models that are based on the data.

2.4.3 Opportunities for Improvement

Smartphone-based GPS travel survey methods provide a mix of passive and self-reported travel data, delivering more accurate OD data and the ability to capture multiple travel days per respondent. These new methods also provide greater coverage of possible OD pairs. Still, the resulting level of OD coverage is much less than what is available from passive Big Data sources, which collect data from a much larger subsample of the population over a longer period of time. That said, smartphone-based GPS travel survey data are more like passive source data than traditional travel diary survey data. As a result, this method offers more opportunities for data fusion by combining household travel survey data and “Big Data.”

3.0 Establishment Survey Data

Traditionally, establishment surveys have targeted people who work at or visit an establishment (employee establishment surveys). Such surveys have been conducted to estimate attraction rates in travel demand forecast models. More recently, with the application of smartphone-based survey methods, commercial vehicle surveys can also be conducted for obtaining information on freight flows.

3.1 Background

In the past, states or planning agencies used employee establishment surveys to estimate trip attraction rates in their travel demand forecast models. The traditional household travel survey did not provide sufficient information to develop trip attraction rates, which is why employee establishment surveys were conducted. New smartphone-based GPS travel surveys often collect as much as seven days' worth of data instead of a single day in traditional household surveys. This yields seven times as much information about nonwork trip attraction patterns, and work trip attraction patterns can also be understood from freely available federal data such as the Longitudinal Employer-Household Dynamics (LEHD) and Census Transportation Planning Products (CTPP) datasets. Passively collected data also now provide cost-effective information on trip attraction patterns and rates.

With smartphone-based methods, the establishment survey targets commercial vehicle movements (commercial establishment surveys) and collects comprehensive data on shipments and vehicles used to move goods. The information gained in these commercial establishment surveys provides significant insight into the origin-destination (OD) patterns to and from sites, rendering them the most relevant type of establishment survey to this body of work. This chapter describes commercial establishment surveys, the processing that is performed to convert the information collected from the survey into an OD dataset, and the resulting uses of this information.

3.2 Introduction

Commercial establishment surveys collect comprehensive data on shipments and vehicles used to move goods—including information on commercial vehicle trips, establishment characteristics, or commercial vehicle fleets. This information can be used to determine employment sizes, industries served, commodity flows to/from the establishment, shipment sizes, distribution channels, and commercial vehicle characteristics. These data can also be used to estimate commodity flow allocation models and mode and shipment-size models and support other freight/commercial vehicle model components.

Several commercial establishment survey methods collect data on site or by phone, mail, online, or smartphone methods. These methods have been used for several types of freight analysis:

- **Commodity flow surveys** are used for goods movement and commercial vehicle analysis, national or statewide freight planning, or modeling. Data collected typically include commercial vehicle trips by mode, origin and destination, vehicle type, commodity type, transfer facilities, and time of day. These surveys can be used to estimate commodity flow allocation models and mode and shipment-size models. They may also support other model components, depending on the availability of the original survey data.

- **Commercial surveys** are used for state or regional freight planning or modeling. Data collected typically include employer characteristics; number and type of employees and vehicles; and aggregate mode, shipment-size, and transfer facility data. These surveys can be used to estimate distribution channel and mode and shipment-size model components of a freight travel demand model.
- **Intercept surveys** are used for corridor or special generator studies, and they can be adapted to be used for freight models. Data are collected at truck stops, weigh-in-motion (WIM) stations, ports, transfer terminals, or facilities of interest. Smartphone-based GPS data collection is another survey approach for these surveys.
- **Truck diary surveys** are used for regional freight modeling. Data collected include commercial trip origin, destination, start time, stop times, routes, distances, vehicle types, commodity types, and stop characteristics. They are the most comprehensive commercial vehicle survey data available and can support estimation of network flow model components—but cannot support upper-level models since they are focused on a single mode. Truck driver surveys can be paired with intercept or establishment surveys to recruit drivers.
- **Vehicle use surveys** are used for vehicle inventory, air quality analysis, and commercial vehicle demand. Data collected include the characteristics of commercial vehicle fleets, which typically include vehicle age, make, model, leasing status, vehicle characteristics, and use. These surveys can be used to estimate commodity flow allocation models and possibly support other model components. They tend to produce small sample sizes for a moderate cost.

One example of an establishment survey in the form of a commodity flow survey is the Bureau of Transportation Statistic's national Commodity Flow Survey (CFS), which is the largest establishment survey in the United States. The CFS data are widely used for freight model calibration and validation. On a local scale, conducting establishment surveys typically includes collecting both outbound and inbound freight movement information, and the correspondence between industries and commodities must be identified to relate outbound and inbound movements. This type of establishment survey is typically conducted by interviewing representatives of specific establishments and gathering information about freight movements in and out of the establishments.

An establishment survey in the form of commercial, intercept, or truck diary survey can provide valuable and detailed information on shipments and an establishment's specifications; however, these are expensive methods for collecting travel data. These surveys gain information on truck travel for different commodity types over different distribution channels with varying numbers of stops. The results can be used to estimate or calibrate freight travel demand models. Commercial surveys generally collect aggregate OD data, while intercept and truck diary surveys generally collect disaggregate OD data. For example, intercept surveys of employees, truck drivers, customers, and delivery personnel collect information on a single trip (OD as well as mode, timing, and purpose) and have been conducted in the New York City, Cincinnati, and Phoenix regions. Further, truck diary surveys collect information on tours as reported daily by truck drivers. Complete daily pattern surveys have been conducted in Calgary, Ohio, Portland, and Colorado. These disaggregate surveys are typically used to estimate truck or freight touring models to simulate urban goods movements on an urban or regional level.

An example of an establishment survey in the form of a vehicle use survey is the US Vehicle Inventory and Use Survey (VIUS). The VIUS surveys a probability sample (stratified by geography and truck characteristics) of all private and commercial trucks registered or licensed in the United States to obtain data on the physical and operational characteristics of the nation's truck population. While it had been collected every 5 years starting in the 1960s, the most recent VIUS was conducted in 2002. California is conducting a statewide VIUS to collect detailed information on typical truck operating characteristics and patterns (such as load factors), vehicle type and attributes, fuel economy, days of operation, and use of marine terminals. The California VIUS will also collect data on commodity types and quantities—including seasonal variation, which may be useful for commodity flow validation. The survey started in early 2016 and will continue through 2017. Data will be collected using a combination of surveys and instrumented vehicles.

3.3 Overview of Data Collection/Processing

3.3.1 Survey Approach

Each of the types of establishment surveys listed above require the following (Cambridge Systematics 2013):

- Defining the geographic boundary of concern.
- Adopting an industry/commodity classification scheme.
- Identifying the universe of companies to survey.
- Determining the sample size.
- Establishing data elements.
- Designing a survey questionnaire.
- Conducting the survey.
- Assembling the database.
- Expanding the data.
- Validating the accuracy of the data.

3.3.2 Survey Questionnaire

The following information is generally developed using an establishment survey:

- General facility information (e.g., name and location).

- Nature of business (NAICS[3]).
- Types of goods/commodities shipped or services provided.
- Quantity of goods/commodities shipped.
- Value of goods/commodities shipped.
- Number and types of vehicles.
- Establishment size and number of employees.
- Specific shipments and their origin, destination, mode, and time of travel.
- Origin and destination of shipments.
- Origins of inbound shipments.
- Destinations for outbound shipments.
- Frequency, size, and weight of shipments.
- Truck stops, locations, and durations to pick up and deliver goods.
- Travel time and cost of travel.

Some surveys may include additional information. These questionnaires can be administered in several ways including via phone, mail, online, or smartphone. Using a smartphone to collect truck movement data is a relatively new method, but one that delivers more data that are of higher quality. To date, this method has not been used for any mode other than truck.

3.3.3 Data Expansion and Validation

Data collection efforts that attempt to use a sample to represent an entire population must also develop a means for expanding the sample data to represent the full population across a chosen market segment (e.g., industry or establishment type). For establishment surveys, this adjustment is usually accomplished by attaching statistical weights to expand the sample data to reflect the characteristics of the entire population of establishments for each segment. Expansion factors can be based on various collected information such as number of employees, amount of output, or size of the establishment. The number of employees is the most commonly used expansion factor.

Before being used, survey responses must be reviewed for accuracy and completeness, and unusable records must be removed from the dataset. Data should be validated by comparing the expanded survey data to other freight data sources with OD information such as the CFS data. Truck count data or roadside intercept survey data may also be used to validate commodity distributions estimated through the establishment survey process. Statistical analysis of the variables collected in the survey process can also help measure result accuracy.

[3] North American Industrial Classification System.

3.4 Data Use in Travel Forecasting Models

Once a comprehensive establishment survey database is assembled, it can be used to develop conventional freight travel demand models and to estimate truck-touring models, advanced supply chain freight models, or urban goods movement models.

3.4.1 Development/Calibration of Conventional Freight Demand Models

One key use of commercial vehicle trip OD information resulting from an establishment survey is to help develop and calibrate conventional freight travel demand forecasting models, which rely on a traditional 3- or 4-step modeling process. For example, after the commercial vehicle trip table is successfully developed from the survey responses, the OD tables can be used to assess the validity of the demand elements of the model. This process is done by comparing the demand for goods and services, the OD patterns, and the modal trips generated by the freight model to the survey data. This comparison should be performed separately for each segment in the freight demand model, typically by commodity group. If mismatches are found between the model output and the survey data, then the analyst must examine each model component and parameter to determine the most appropriate adjustments.

3.4.2 Development/Calibration of Truck-Touring Models

Establishment surveys can be used to estimate truck or freight touring models that simulate goods movements on statewide, urban, or regional levels. Establishment commercial vehicle survey data are used to estimate parameters for tour patterns, stop duration and location, and timing (among others). These surveys are also used to calibrate an existing truck-touring model that is transferred from another geographic area. This method is appropriate when the surveys do not contain all the parameters required by the model—calibration can focus on the parameters that are available in the local survey data.

3.4.3 Development/Calibration of Supply Chain Models

Establishment surveys that collect data on commodity flows for long-haul, multimodal freight movements can be used to estimate or calibrate supply chain freight models that simulate goods movement on a national level. Establishment commodity flow survey data are used to estimate parameters for supplier selection, distribution channels, and mode and shipment sizes (among others). These surveys are also used to calibrate an existing supply chain model in a geographic region, such as a state or megaregion. This is appropriate when the long-haul goods movements are needed to support megaregion, state, or regional freight analysis.

3.4.4 Other Planning Analyses

Establishment survey data also have considerable utility for nonmodel planning analyses. Freight planning agencies often use commodity flows, vehicle characteristics, and vehicle trips characteristics information resulting from establishment surveys as part of freight planning, mobility, and economic analyses.

3.5 Assessment of Data Suitability

This section discusses the strengths and shortcomings of establishment survey data and improvement opportunities.

3.5.1 Strengths

The fundamental strength of establishment survey data and the resulting OD flows is that these data represent commercial vehicle movements and behaviors that illustrate the movement of goods and services across the United States. This information is useful for state and regional planning and modeling purposes. Mobility for goods and services is an important economic driver for the United States, and establishment surveys are critically important to understanding and planning for transportation investments to support this mobility. Establishment surveys that collect commodity flow data can be used to estimate supply chain models. These models provide an understanding of the economic drivers for freight movement and the parameters affecting mode choice to support federal, state, and regional planning activities. Establishment surveys that are primarily commercial surveys have the benefit of being less expensive than truck diary surveys. They can be used for model calibration or validation if OD patterns by commodity are collected and expansion of the surveys is completed. Establishment surveys that include truck diary components can be used effectively to estimate model parameters for behavioral truck-touring models. These are new advanced behavioral/agent-based freight models being developed to support state and regional policy needs and issues. Vehicle inventory and use surveys provide an understanding of payload factors and vehicle types and are most useful in translating commodity flow data into truck assignments. These surveys are uncommon, as most agencies relied on the US VIUS for these data. As mentioned previously, the last US VIUS was conducted in 2002.

3.5.2 Shortcomings

An establishment survey can provide valuable and detailed information on shipments and an establishment's specifications; however, in general, it is an expensive method to collect freight movement data. Establishment surveys that collect either OD data or full driver diary data can be nearly as resource-intensive as household surveys, primarily because of recruitment costs. This aspect is a significant challenge of truck diary, vehicle use, and commodity flow surveys. Survey costs vary widely depending on the type of commercial vehicle survey being conducted, the desired sample size, and the type of data required (e.g., $350k for the Portland region[4] to $1.5 million for the Phoenix region[5]). Commercial and intercept surveys are usually less expensive and require fewer resources compared to truck diary, vehicle use, and commodity flow surveys.

The other significant challenge of establishment surveys is low response rates. Most of the establishment surveys conducted recently in the United States have struggled with low response rates and small sample sizes with excessive costs. In most cases, recruiting participants for establishment surveys is challenging because many establishments are concerned about drivers

[4] Portland Metro Establishment Survey, 2016 (collected data on ~28,800 drivers/trucks and ~723,000 trips).

[5] Maricopa Association of Governments Commercial Establishment Survey, 2016 (to collect data at 416 establishments in the MAG region).

being distracted, confidential data being shared, or respondents taking time during their work day to complete the survey. The difficulty in recruiting participants is primarily a concern for state, regional, or local surveys. National-level surveys, on the other hand, have much higher response rates because: 1) respondents are confident that the data remain confidential; and 2) the federal government can require responses. Generally, intercept surveys and national surveys have higher response rates compared to other establishment survey types. In addition, the data collection methods are undergoing tremendous changes due to new survey technologies (such as smartphones) to reduce burden, increase response rates, and improve accuracy.

3.5.3 Opportunities for Improvement

Establishment surveys can benefit from improvements to the survey design and data collection process. For example, questionnaires' designs could prioritize simplicity rather than aiming for thoroughness and gathering unnecessary information. As discussed, low response rates are a significant shortcoming of these types of surveys. Short and efficient surveys are more likely to have higher response rates. Also, providing incentives in the establishment selection process to encourage establishments to respond would likely improve the response rates and is another improvement opportunity. In addition, agencies could rely more on intercept and national surveys in which response rates are higher. Efficient questionnaire design can be achieved by having focus groups and stakeholders from selected establishments review the data elements and the entire survey questionnaire. Feedback from the focus groups can then be used to design, revise, or improve the survey.

Establishment surveys that collect either OD data or full driver diary data can be resource-intensive. To address this issue, partnerships with other agencies within the region could provide an excellent opportunity for shared data collection for freight. Co-funding large establishment survey efforts can significantly decrease the high costs of surveys. Truck diary, vehicle use, and commodity flow surveys can benefit from this approach. This technique was implemented in Arizona where the Maricopa Association of Governments (Phoenix) and the Pima Association of Governments (Tucson) have shared data and modeling resources for passenger, freight, and land-use modeling systems. Another example of this approach is in Texas, where the Texas Department of Transportation conducted truck diary surveys for five metropolitan planning organizations (MPOs) across the state, and the joint effort helped to increase sample size beyond a single MPO.

New technologies offer an alternative/newer method for commercial vehicle surveys by using GPS technology in trucks to collect commercial vehicle trip data. Commercial trip origins, destinations, start times, stop times, routes, and distances can be collected passively with this approach. However, truck GPS data have some limitations. These data lack crucial data variables (e.g., commodity type or establishment characteristics) and are difficult to expand, which is required to improve comprehensiveness and representativeness.

Smartphones can also be used to collect additional data on vehicle type, shipment weights, commodities carried, parking or tolls paid, and stop purposes. Smartphones capture and precisely route all trips. Intercept, truck diary, and commodity flow surveys can benefit from smartphones-based GPS data collection, which can decrease the percentage of incomplete surveys compared to paper. It can also reduce respondent burden and increase the response rate.

4.0 External Survey Data

4.1 Background

External travel surveys are a specialized type of travel survey focused on people who regularly travel into, out of, or through a region or study area. In many cases, long-distance or interregional commuters tend to dominate the survey participants. External travel is unique because it uses transportation infrastructure in a region, but the traveler may live outside the region and thus not have a direct voice in the decision-making process regarding transportation in the region. In many regions, external travelers generate a disproportionate demand on the transportation system. A recent study in the San Francisco Bay Area found that nearly 15% of the vehicle miles traveled in the region came from interregional trips, but that less than 5% of the trips were interregional (Cambridge Systematics 2017). External travel may also be indicative of other planning issues such as housing affordability or jobs-housing imbalance. In these situations, targeted external surveys inform planners and policy-makers about the travel behavior of visitors, and the results can guide infrastructure and policy decisions that may have been overlooked in a traditional household travel survey.

4.2 Introduction

External surveys are used to collect traveler information and behavior from people who travel into, out of, and through the region. The data gathered from these surveys inform travel models about the scale and characteristics of travel to and from external cordon zones. The key differentiator between an external survey and a visitor survey is the residence location and length of stay of the visitor. External surveys typically focus on travelers entering or exiting the region on the same day; interregional commuters are the most common example of an external traveler. Other common external trip purposes include shopping trips, doctor visits, and personal trips (e.g., visiting family and friends). Visitor surveys tend to focus on overnight visitors.

4.3 Overview of Data Collection

4.3.1 Survey Universe Identification

The survey universe for external travel surveys is usually defined as travelers who enter and exit the region on the same day. In most cases, overnight visitors are excluded from this survey universe because their travel behavior is often different than travelers who enter and exit the region on the same day and would typically require a different survey instrument. The sample rate is typically obtained by analyzing traffic counts at major roadways crossing the boundaries of the region (also known as cordon locations). In regions with large rail or bus interregional commuters, the universe may also include external transit users. Designing the sample based on traffic counts and transit boardings provides the ability to expand the data to those counts after data collection.

4.3.2 Survey Questionnaire

The survey's primary purpose is to gain insight into the magnitude, purpose, origin\destination, timing, and mode\occupancy of trips, along with information about the socioeconomic characteristics of the traveler. Typical external surveys ask travelers to provide the following types of information.

- Trip characteristics:
    - Origin and destination of the trip.
    - Travel mode(s) of the trip.
    - If transit is used, the access and egress mode and boarding and alighting station.
    - Time of day.
    - Duration of the trip.
    - Route, if possible.
    - Number of and relationship among participants.
    - Purpose.
    - Frequency of trips made into region.
    - Number and location of stops made by the traveler in the region.
- Socioeconomic characteristics:
    - Residence location of traveler.
    - Income.
    - Ethnicity.
    - Household size.

Depending on the goals of the survey, it may also be useful to ask respondents why they are traveling into the region. This information can help planners and policy-makers better understand the motivation to travel into the region. For example, it might be useful to determine why regular travelers into the region do not live in the region. The traditional travel survey would reveal the behavioral choices, but it would be difficult to accurately deduce the justification for the behavior without a more formal survey such as a stated preference survey.

4.3.3 Recruitment Methods

4.3.3.1 *License Plate Intercept*

License plate intercepts are typically used to recruit respondents since in-person interviewing can be costly and may be prohibited by law. With this type of recruitment, a license plate reader is placed along the roadway at the external station. The license plate numbers are provided to the department of motor vehicles to match address and other contact details. The addresses are then used to either send the survey instrument to respondents or inform potential respondents that they may be contacted by telephone, in which case the sample is then contacted by phone and information is gathered using prompted recall. In some areas, due to privacy concerns, it may not be possible to get specific addresses of travelers with this method. However, even more coarse information about the driver such as a ZIP Code can help target recruitment in specific communities.

4.3.3.2 *Random-Digit Dialing and Address-Based Sampling*

Random-digit dialing or address-based sampling methods are used to recruit respondents when a concentration of home or work locations of external travelers is known. Commuter home location, for example, is available in the U.S. Census Longitudinal Employment and Housing Dynamics (LEHD) program. In this situation, the recruiter randomly selects phone numbers or addresses in a geographic area where known external travelers reside or work to recruit potential respondents. The recruiter will usually first ask whether the person answering meets the definition of an external traveler, and, if so, whether the person would be willing to enter the survey. However, with increasing use of Voice over Internet Protocol and number portability, geographic isolation via telephone is becoming more difficult. These methods also have documented response biases that would also need to be mitigated through other recruitment means.

4.3.3.3 *In-Person Intercept*

In an external survey, survey participants can be intercepted in person at or near an external boundary or cordon location. For example, survey participants could be approached and recruited at a commuter rail or bus station as they are waiting to board a transit vehicle. In regions where interregional travel is primarily by automobile, in-person intercepts in the roadway are impractical and may be illegal.[6] In an Omaha external survey, Metropolitan Area Planning Agency staff intercepted individuals at rest areas, truck stops, and gas stations near the external cordon. (Farnsworth & Hard 2013)

4.3.3.4 *Advertising*

In many regions, external travelers often enter the region via a private vehicle. In this case, advertised recruitment could be done with billboards along primary entry and exit routes or radio advertisements during drive-time commutes. These traditional media approaches require a user to remember the website or phone number and engage the survey at a more opportune time, which hinders recruitment effectiveness. In fact, the Federal Highway Administration (FHWA) advises it is unsafe to request motorists to write down telephone numbers or websites while they are moving. FHWA and research indicates directing potential respondents to simple phone numbers like 311 and 511 may improve recall among motorists and be safer. In some cases, a larger public education campaign could be conducted with a kickoff press conference and other community outreach to solicit participation.

Social media platforms also offer an opportunity to target advertisements to certain geographic areas and sociodemographic characteristics. With these newer, more targeted recruitments, links can be provided to users to click-through, increasing recruitment effectiveness.

[6] In 1994, the Kentucky Attorney General and Indiana Department of Transportation issued an opinion stating that roadway intercept surveys were illegal under the Fourth Amendment, because travel information could be obtained through less obtrusive means. Legal Opinion from the Kentucky Attorney General on Legality of Highway Use Survey by Kentuckiana Regional Planning and Development Agency.

4.3.4 Collection Methods

4.3.4.1 *Computer-Assisted Interview or Online Recall*

This method uses an interview to assist survey respondents in recalling the details of their travel. The interview can be done face-to-face or with the interviewer via phone. In some instances, the interview is conducted completely online using a web-based survey without the assistance of a human interviewer. The interview often covers travel that occurred on the current day or the previous day. Recall bias becomes more significant for each preceding day the survey attempts to recall. This survey uses a computer program to prompt questions for the respondents. Like smartphone survey data collection, the respondent's answers are checked in real time allowing the surveyor to follow up with additional questions to address errors or inconsistencies. This method, like the smartphone-based GPS travel survey method, also allows for the survey to branch and lead toward other questions (e.g., fare questions for transit trips but not walking trips), when warranted.

4.3.4.2 *Smartphone-based GPS Travel Survey*

Smartphone-based GPS travel survey methods offer the most accurate method for collecting travel survey data. Once respondents are recruited, they are prompted to download an app from the appropriate store for their mobile device's operating system. The app, once enabled by the user, will passively track the movements of the respondents, and it will prompt the user for travel details once it senses the trip is completed. This platform provides much more accurate origin, destination, and time of travel information than computer-assisted interview and prompted recall methods and has the cleanest and most complete data. This reduces the recall bias for this method more than other traditional approaches.

4.3.4.3 *Paper-based Recall*

The paper-based recall method uses a travel survey printed on a piece of paper. Respondents are recruited to complete the survey on their own or with the assistance of an interviewer. The respondents are asked questions about their travel behavior on the current day or the previous day. The respondent's answers are recorded on a piece of paper and later entered into a computer database. These forms can be distributed in large quantities by a small number of staff, reducing the expense of fielding a survey. This type of survey collection is the cheapest and easiest to administer, but many surveys will be rejected during the data validation phase because of incomplete or inconsistent responses. While being subject to the same recall bias of computer-based recall surveys, the paper-based recall offers few opportunities to intervene with respondents if their responses include inconsistencies or errors.

4.3.4.4 *Bluetooth Detection Methods*

Through trips, or external-to-external trips, can be measured using Bluetooth devices. Bluetooth devices are placed at external cordon stations and can be used to estimate the percentage of "through" traffic, or traffic entering\exiting at each external station pair. These data can then be used as another control in the data expansion process, providing a more precise estimate of internal-external, external-internal, and external-external flows. Bluetooth detectors can also be placed around key destinations inside the region to help identify the magnitude of travel between origin-destination (OD) pairs from outside the region to specific areas within the region. This type of analysis will further assist with scaling results after the surveys are complete.

4.4 Assessment of External Survey-based OD Matrix Suitability for Travel Forecasting

4.4.1 Strengths

External travel surveys offer a comprehensive view of travel behavior into a region or study area. This nonresident travel is not captured by household surveys, and the frequency of external travel for residents is typically too low to adequately describe this important travel market for modeling. External surveys can be used to construct origin-destination matrices for travel forecasting. A strength of some of the external survey methods is that they include characteristics of the traveler and the trip, information that is not available using passively collected data.

4.4.2 Shortcomings

The following issues require consideration:

1. External travelers are difficult to identify and solicit survey responses from. The most straightforward capture method using a traffic stop at external cordon points raises safety concerns, causes delay, and, may be illegal in some areas due to privacy invasion. These surveys often need to be paired with other recruitment methods to generate significant responses.
2. OD matrices suffer from low sample rates. The small number of surveys collected and trips observed often offer only a sparse matrix of origins and destinations at a detailed geographic level, especially when segmented by purpose, time-of-day, mode\occupancy, and socioeconomic segment. Care must be taken to ensure that there are enough samples in each segment to minimize survey bias. As discussed, Bluetooth detectors could also be used to improve the expansion process, but this does not address problems with low sample rates.
3. The low sample rates mentioned above often limit the ability of the analyst to include potentially useful market segmentation variables in external travel. For example, the time of day of travel is typically only available in a few broad time periods. This limits the ability of the models to represent the effects of congestion and other policies that vary by time of day on external travel markets.

4.4.3 Opportunities for Improvement

The introduction of new passively collected, Big Data sources over the last decade offers new opportunities to refine and improve external survey implementation. These additional data sources from mobile carriers, navigation devices, and location-based smartphone apps can help more effectively target the surveys to regular external travelers. These sources can also be combined with traditional survey methods to provide a richer view of external-internal and internal-external travel within regions, including more detailed OD information.

With advances in location-based apps and services, external travel information can now be derived in large part from private data vendors. Data from location-based apps offer an opportunity to build large, disaggregated travel datasets with attributes like trip purpose, mode, time of day, route, and certain socioeconomic characteristics. These data sources can be mined to identify external travel into a region or study area. The passively obtained external travel survey can also be remined if the boundaries of the study area change, eliminating the need for multiple field efforts.

5.0 Intercept Survey Data

5.1 Background

Intercept sampling is a method to recruit individuals into a survey. A respondent is selected at random—or “intercepted”—during an activity of interest. This recruitment method is reliable because people are selected while engaged in a travel behavior of interest to the survey (e.g., while riding public transit). In transportation planning, intercept surveys are a convenient way to identify and survey special or small markets of interest. The most common type of intercept recruiting in travel surveying is the transit on-board survey. In transit on-board surveys, respondents are selected at random on a transit vehicle or in a transit station. Other commonly surveyed markets include toll, bike, visitor, external travel, and airport surveys.

5.2 Overview of Data Collection

Survey participants can be intercepted in person at or near a location of interest. In-person intercept surveys tend to benefit from higher sample rates when respondents are stationary rather than moving. For this reason, most transit surveys are conducted as passengers are riding transit vehicles rather than as they are waiting to board the transit vehicle. In an Omaha external survey, Metropolitan Area Planning Agency intercepted individuals at rest areas, truck stops, and gas stations near the external cordon rather than stopping cars at the cordon. (Farnsworth & Hard 2013)

Intercept survey data must be expanded to total estimates of travelers depending on the type of travel market being intercepted. For example, transit on-board surveys are typically expanded to system-wide transit boardings. Air passengers surveyed for development of an airport ground access model are expanded to estimates of enplaned (nontransferring) passengers, and so on.

5.3 Assessment of Data Suitability

5.3.1 Strengths

Intercept surveys are preferred in specialized travel markets because potential respondents are relatively easy to identify and recruit compared to other sampling methods such as random-digit dialing or address-based sampling approaches. To build a valid model, travel surveys should have enough completed surveys across several travel markets to build statistical inferences about travel choices. For example, in mode choice estimation, it might be useful to examine gender, age, and income as explanatory variables in the model. This requires having sufficient survey responses in each crosstabulation to make statistical inferences about mode choice behavior. In a household survey that only collects a few hundred transit or bicycle trips, it is unlikely that the survey will yield enough valid surveys in each submarket tabulation to build a statistical inference about behavior. Intercepts offer an effective way to isolate specialized travel markets to augment household surveys.

5.3.2 Shortcomings

Intercept surveys are prone to several biases that can affect survey results. When a sampling plan is being developed, the survey team should be aware of the following biases and adjust the sampling plan according to local conditions.

1. Participation Bias: As with many surveys, intercept surveys distract people from what they would otherwise be doing. In an intercept survey, people may be more likely to spend the time completing the survey if they are engaged with the topic—either positively or negatively. These individuals are more likely to answer questions in a way that conveys strong feelings, while those with more moderate feelings may not participate.
2. Recruitment Bias: Recruitment bias is when the recruiter either consciously or subconsciously systematically excludes a class of people from the survey. An example of this type of bias may be an English-speaking recruiter passing up opportunities to recruit Spanish-speakers due to language barriers. In this case, the survey would be biased toward the opinion of non-Spanish-speaking residents. All survey methods must manage for recruitment bias.
3. Recall Bias: Intercept surveys in transportation often ask people about their travel over the preceding day or week. Most people will not accurately remember their trip-making behavior the further in the past it was. Survey participants are most likely to forget short and nonmotorized trips because they were not deemed important. This bias can be minimized by prompting individuals about their trips during the survey, using online validation tools during the survey, and limiting the recall to the most recent few hours.

6.0 Longitudinal Employer-Household Dynamics Data

6.1 Background

Another specialized type of origin-destination (OD) travel dataset can contain information on a single purpose of travel rather than all person trips. One notable example of purpose-specific travel data is origin-destination datasets on the journey-to-work or work commute. This chapter describes one such dataset, the Longitudinal Employer-Household Dynamics Data (LEHD), while another such dataset, the Census Transportation Planning Package (CTPP) is described in Chapter 7.0

6.2 Introduction

The work commute has been a topic of special interest to transportation planners for several reasons. The work commute (including stops along the way) accounts for roughly one-third of trips in most urban settings and an even somewhat higher proportion of vehicle miles traveled (because work trips tend to be longer, on average). Moreover, because a large portion of work trips are concentrated in peak periods, these trips are both a major cause and victim of peak-period congestion. Work commutes also have a disproportionate effect on public transit, as travelers are more likely to use public transit for their work commute than for other purposes. Finally, work trips are often viewed by planners as having a higher economic value than many other purposes as they supply jobs to citizens and labor to companies, thus supporting a vital function in the economy.

The LEHD program is administered by the Center for Economic Studies at the U.S. Census Bureau. It involves a voluntary partnership of state employment security agencies (ESAs), the federal Office of Personnel Management (OPM) and U.S. Bureau of Labor Statistics (BLS) with the U.S. Census Bureau. Through this partnership, states agree to share their Unemployment Insurance (UI) earnings data and Quarterly Census of Employment and Wages (QCEW) data with the U.S. Census Bureau in return for the processing of these data together with the bureau's own data to produce statistics on employment, earnings, and work commute flows at detailed levels of geography and industry for different demographic groups. The program produces these data entirely from the combination of existing datasets, avoiding the burden of additional data collection.

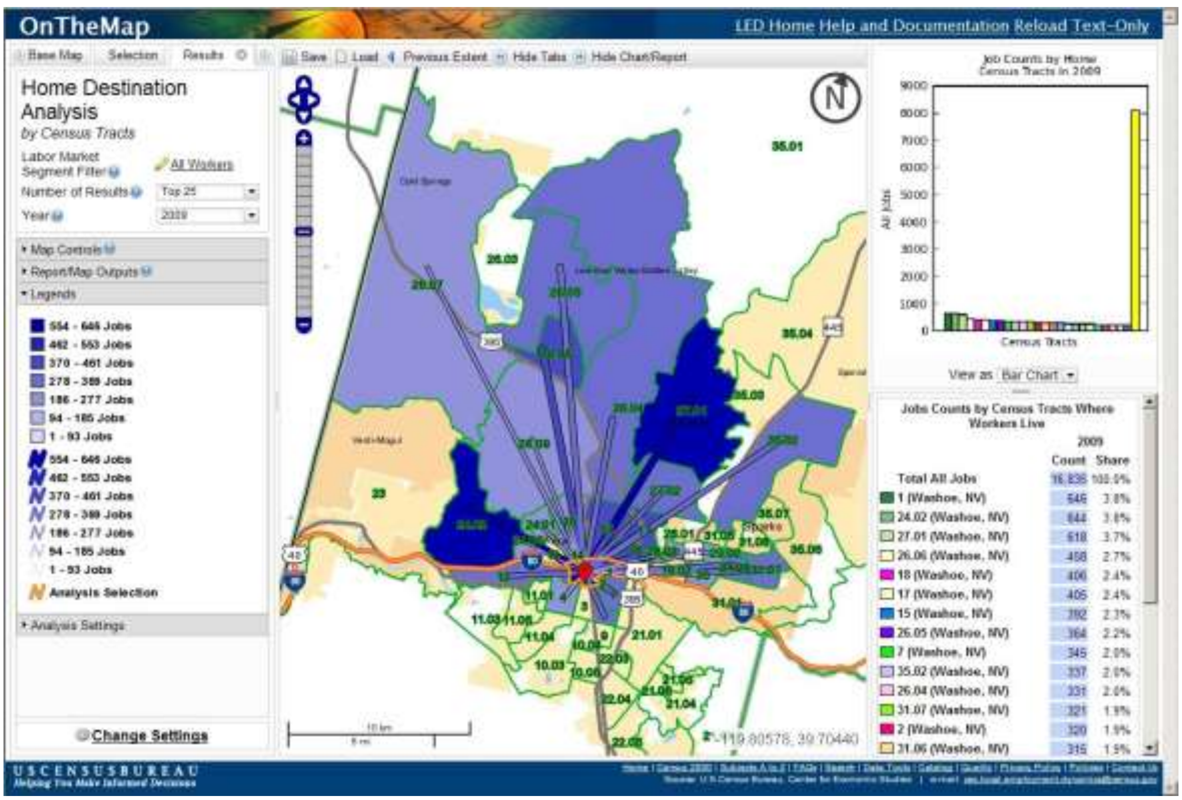
This approach, using administrative records such as the UI and QCEW data, distinguished the LEHD data from similar data on work commute patterns from the ACS/CTPP, the National Household Travel Survey (NHTS), or other travel surveys. The data are based on complete coverage of jobs/workers (covered by UI) rather than a sample (such as the approximately 7% sample on which the ACS/CTTP data are based), and thus involves the processing of data on every individual worker and company in the participating states. Both the secondary use of data collected for other purposes, and the volume and variety of the data involved, make it reasonable to classify the LEHD data as Big Data.

Like other Big Data sources of OD data, LEHD provides a much more complete dataset in the spatial dimension of travel, especially at finer levels of geography, and the LEHD contains many more OD pairs than the CTPP. It is not a sample—it is data on all employers and all employees (covered by unemployment insurance or the equivalent for civil federal government employees). In general, this is a significant advantage over sparse, survey-sample-based datasets; however, there are some important concerns that—in the case of the LEHD data—at least some of the additional OD pairs are inaccurate.

The following sections present the content and coverage of the LEHD data, the sources and creation of the LEHD dataset, its use in travel modeling and forecasting, and an assessment of its strengths, shortcomings, and opportunities for improved use.

6.3 Data Contents and Coverage

The LEHD “OnTheMap” website[7] provides several tools for reporting, mapping, and downloading data (Figure 5). For many travel modeling and forecasting applications, the user is most interested in downloading the data, but the visualization features of the website can be helpful for exploring the data and in performing basic data validation for the region and time period of interest.



Source: https://onthemap.ces.census.gov/.

Figure 5. Screenshot of LEHD OnTheMap website visualization.

The LEHD data products include three types of data describing employment or jobs, workers (by place of residence), and the OD flows from workers’ residences to their jobs. These three data products are presented and referred to as the following:

[7] U.S. Census Bureau, On the Map

- Residence Area Characteristics.
- Workplace Area Characteristics (WAC).
- LEHD Origin-Destination Employment Statistics (LODES).

The program is also beta-testing a worker relocation dataset as of October 2017. This discussion primarily focuses on the LODES dataset and, secondarily, the WAC dataset, which is commonly used with it in the context of travel modeling and forecasting. These datasets are available for download at the census block level and can be aggregated to any level of census geography. These datasets can also be aggregated to many models' traffic analysis zone (TAZ) systems if they conform to block boundaries (which many, but not all, do). Workers can be segmented by industry, age, income, gender, race, ethnicity, and education. Firms can also be segmented by size (number of employees) and age of the firm.

Since the program is based on a voluntary partnership with the states, the data are not available for all states in all years, but only when states have participated in the partnership. Although data are available back to 2002 for some states, complete data for all 50 states is only available from 2011 to 2013; Wyoming has not participated since 2013, but the other 49 states have since participated together with several of the US territories.

The employment (jobs) included in the LEHD data are limited to those covered by UI. This includes most jobs, including both primary or full-time, as well as part-time and secondary jobs; this contrasts with CTPP, which is only concerned with a worker's *primary* job. The main group not covered by UI is sole proprietors who are generally estimated to account for between 5% and 20% of all jobs in the United States. Since many sole proprietors work at home or establish a place of business nearby and convenient to their homes, the lack of sole proprietors in the LEHD data may be one reason why short-distance commutes are underrepresented.

Other groups exempt from UI include federal employees, including the military, and railroad employees. These usually account for anywhere from 0% to 20% of the workforce of most regions. However, the partnership with the OPM allows most federal civil employees to be included in the LEHD data, but railroad and military employees are still excluded (and a few other small groups such as employees of the Nuclear Regulatory Commission, among others).

6.4 Data Sources and Creation

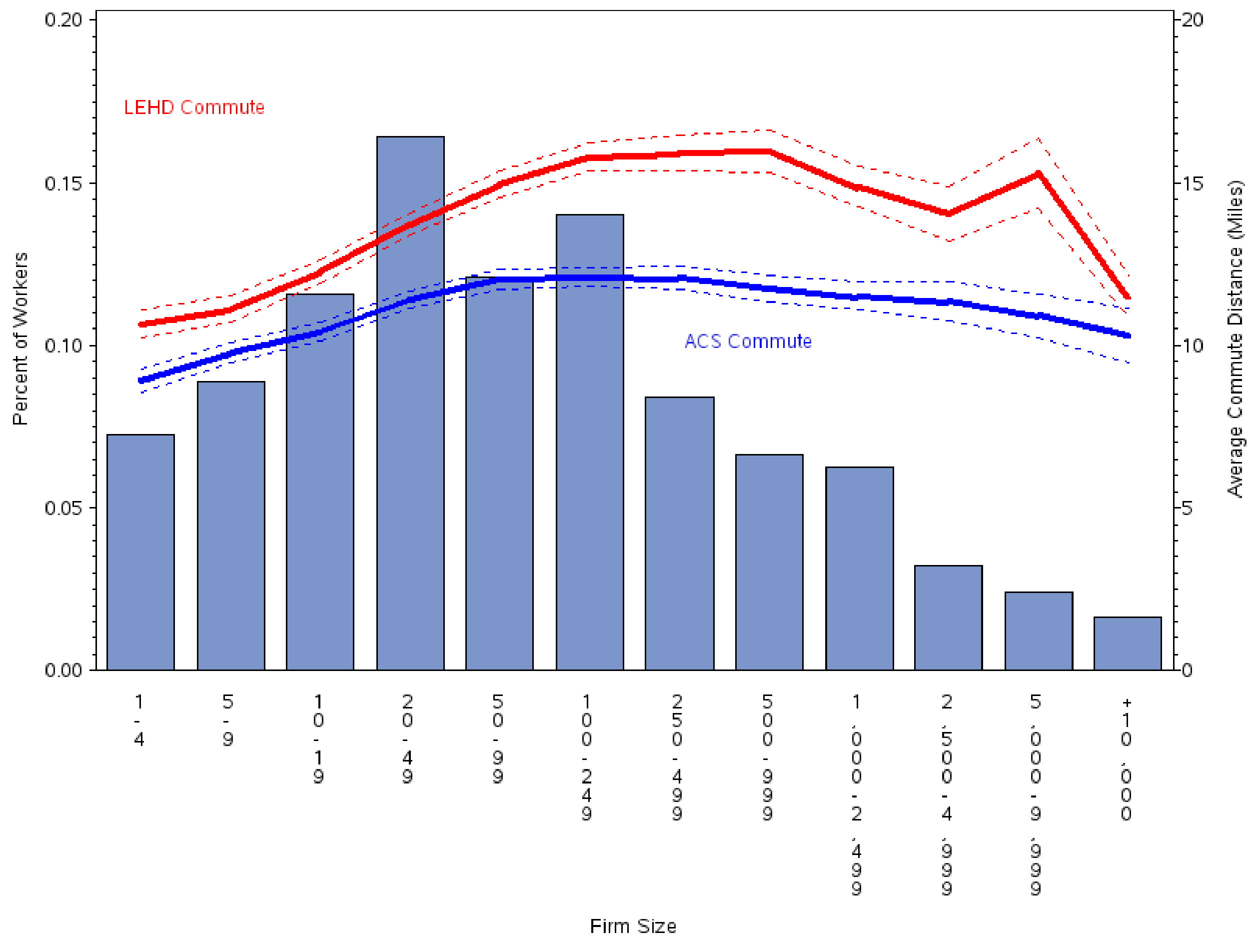
The LEHD datasets and LODES are created through a process of data fusion, primarily of administrative records. Workers' Social Security numbers (SSNs) are integral to joining various source datasets to provide the LEHD data with work residence-to-workplace flows segmented by age, gender, income, and firm size, among.

The following administrative datasets provide the source data for the LEHD datasets:

- The **QCEW** collects quarterly information on jobs, wages, and establishments throughout the country that are covered by state UI programs, and federal workers covered by the Unemployment Compensation for Federal Employees program. QCEW is managed by the BLS and was formerly commonly referred to as ES-202 program/data.

- The **Personal Characteristics File (PCF)** is a table based primarily on the information supplied in the application for a SSN from the Social Security Administration and contained in their Numerical Identification System ("Numident") file, which serves as the basis of the PCF. This provides information including a worker's gender, date of birth, race, and citizenship, among others.
- The **Composite Person Record** is a data table derived from multiple sources including the Internal Revenue Service, Medicare, and the U.S. Department of Housing and Urban Development; the SSN is used as a key to gain information on the worker's current place of residence.
- **Wage Records** from the state ESAs provide a table of employees by SSN linked to their employer.
- **Multiple Worksite Report (MWR)** reports workplace information for employers with multiple worksites. They are required in roughly half of the states and requested in the others, although compliance is not perfect even in states requiring it and much lower in those that do not. Without an MWR, all employees of a firm may be shown at the employer's primary address. This "headquartering problem" is one of the key limitations of the LEHD dataset. Various studies have estimated that it affects roughly 45% of jobs in the LEHD nationwide. The effect, however, is not uniform, with much greater effect in states without mandatory MWR reporting. Local government is known to be among the least compliant industries in filing MWRs with many employees (e.g., teachers) reported at a single location (e.g., the school district offices). Moreover, the problem is at the level of states because this is the level of reporting, so employees of companies with worksites in multiple cities in the same state may report all employees at just one location that may be hundreds of miles away.

The LEHD program attempts to reduce the issues resulting from the headquartering problem by using an imputation known as the Unit-to-Worker (or U2W) method based on a simple gravity model calibrated based on Minnesota data (because it was originally the only state with complete data), to distribute workers among their employers' workplaces based on distance to their residences. Recent research by Green et al. (2017), however, suggests both that this is the largest source of the discrepancy between LEHD and CTPP trip-length distributions (Figure 6). This research also suggests that the simpler assumption—that the worker works at his or her employer's closest location—is substantially better.



Source: U.S. Census Bureau (Green et al. 2017).

Figure 6. Average trip length in LEHD and ACS/CTPP, by firm size.

Before the data are finalized and used as the LEHD dataset, however, a process of **disclosure proofing** is applied. In this process, noise is intentionally introduced into the data at low levels of geography to protect the confidentiality of workers and firms. The process is not well documented, but it is understood to focus on obscuring employer and worker characteristics rather than flows and believed to be accomplished by swapping workers or firms in different blocks within the same unit of more aggregate geography. This is done so that the data integrity is protected at that higher level of geography.

6.5 Data Use for Travel Modeling and Forecasting

The LEHD data are commonly used in travel models in two ways and for two purposes. The WAC dataset is commonly used either to estimate employment data for a model's zone system or to validate employment data from some other source. The LODES dataset is commonly used to validate the model's distribution of work trips/commutes. In traditional, 4-step models this would be the home-based work gravity model; in more advanced models this would generally be a (destination) choice model for the workplace location. This validation is sometimes limited to examination and comparison of trip-length frequency distributions; however, this is inadvisable on multiple grounds. In this case, it is inadvisable primarily because the LEHD trip-length frequency distribution is known to be skewed relative to other (survey-based) data, largely due to its headquartering problem. Better validation techniques using LEHD include using goodness-of-fit statistics or aggregate district-to-district flows. LEHD can also be used for incremental forecasting, primarily by using a model's forecast to pivot off the LEHD's current-year distribution or to estimate gravity or destination choice models. However, both techniques are generally believed to be rarely used in practice.

6.6 Assessment of Data Suitability

The LEHD data have great potential. However, issues such as the headquartering problem and, to a lesser extent, the disclosure proofing have limited its usefulness and use up to this point. However, with some of the insights afforded by recent research, it may be possible to better correct for these issues and make better use of the data in the future.

6.6.1 Strengths

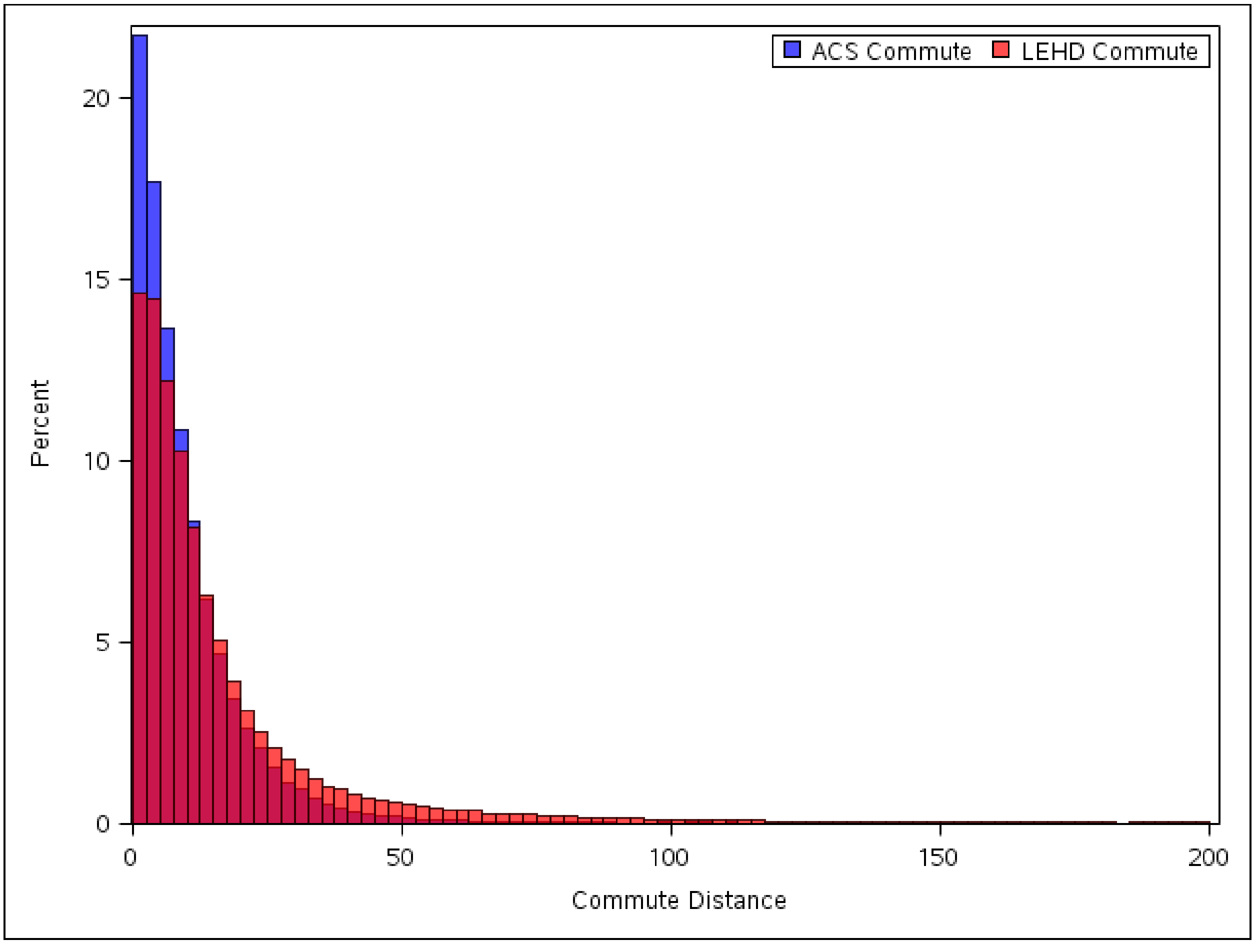
One of the greatest strengths of the LEHD data is that these data are freely available (for participating states). Obviously and rightly, this greatly lowers the barrier to its use. Secondly, the LEHD data's resolution at the census block level is a significant advantage. Many sources of employment data, including other reported forms of the QCEW data, are only available at the county level, which greatly limits their usefulness for modeling at the urban scale. Moreover, while CTPP data are available for census TAZs, current model TAZs frequently differ from census TAZs for several reasons. These reasons include initial differences between the two schemes due to rules related to delineating census TAZs and subsequent revisions of model zone systems.

Another key strength of the LEHD data, and one that is commonly overlooked or underappreciated, is its derivation from a complete dataset rather than a small survey sample. Because of this, it offers far more complete coverage of the OD space and does so at a much finer level of resolution. The completeness of the underlying source data gives the LEHD the potential to be a tremendously powerful dataset.

6.6.2 Shortcomings

The key limitation of the LEHD data is that these data have been observed to be significantly and systematically biased regarding trip length. Comparisons with both the census's own ACS/CTPP data (shown in Figure 7) and the NHTS and various household travel surveys conducted via computer-aided telephone interviewing, online, and smartphone, all show substantially more short-distance and fewer long-distance commute trips.[8] The skewed trip lengths, and OD patterns more generally, have limited both the use and usefulness of the LODES data; however, several simple approaches can significantly improve the representativeness of the data, as discussed in the following section.

[8] Interestingly, this trip length bias is similar to that observed in other forms of Big Data collected passively from mobile and in-vehicle devices, although the underlying mechanism is understood to be different. While the bias in passive data is due to the sampling probability being linked to duration, in the LEHD, the primary cause of the bias appears to be the headquartering problem and its handling.



Source: U.S. Census Bureau (Green et al. 2017).

Figure 7. Comparison of work trip-length frequencies from LEHD and CTPP.

The disclosure proofing has also limited the use of the data. Although the data perturbations are generally limited to a relatively small scale geographically, they can significantly limit the ability to compare different vintages of the data to understand changes in employment and commuting patterns. Another shortcoming of the LEHD data, as addressed above in this section, is the headquartering problem. Analysts using LEHD data should be aware of the potential overcounting of jobs associated with headquarters locations and adjust for overcounting, as appropriate. Lastly, the lack of information on mode and departure time also limits the data as many analyses require information on peak-hour commuting and transit commuting, among others. However, the nature of the source data precludes the possibility of including these attributes.

6.6.3 Opportunities for Improvement

Using LEHD for transportation modeling and planning has been limited by the aforementioned data quality issues related to systematic biases, several practical approaches can and have been used to limit and correct for these issues, and insights from recent research may lead to improved versions of LEHD in the future. One of the simplest approaches is to limit use of the data to commute trips internal to an urban area. Most of the inaccuracies are related to trips longer than 50 or 100 miles. Hence, excluding commutes extending beyond an urban model area can exclude many of the worst inaccuracies. This approach is widespread and commonly used to improve the LEHD data. In contrast, LEHD (in its current form) should generally not be used to understand external commute patterns. As with other traditional OD data sources, it is important to compare datasets and validate the data for use in a model.

The underreporting of short commutes can also be an issue within metropolitan areas. This issue can largely, but not entirely be corrected using a second technique in which an alternative source of employment data is used to rescale the LEHD trips at a zonal level. Much of the deficiency is related to companies underreporting employees at “branch” business locations. If a second source of employment data is available in which the headquartering problem has been corrected, then it can be used to identify and largely correct the improperly assigned trips in the LEHD data. This approach is not as common, but has been applied, as in Chattanooga, Tennessee. Given the new insights from recent research, the U.S. Census Bureau may revise and substantially improve new vintages of the LEHD data. For example, the replacement of the U2W imputation algorithm could lead to substantially less biased data. Were this approach to be used in combination with the rescaling described above, it could produce substantially better data, potentially better than current alternatives. The omission/exclusion of sole proprietors may still lead to some underreporting of short commutes, which may be important to consider for nonmotorized modeling, but even with this issue, LEHD rescaled to address headquartering issues may be one of the best sources of data on commuting patterns.

7.0 Census Transportation Planning Package Data

7.1 Background

Beginning with the 1970 census, the U.S. Census Bureau began asking questions about journey-to-work (JTW) (or commuting) characteristics. The questions in the decennial census inquired about mode of transportation to work, time spent traveling to work, and workplace location. When combined with other census information about income, auto ownership, and household size, these data proved useful to transportation planners and policy-makers. Since 1970, the U.S. Census Bureau and its partners have synthesized census data into a transportation-specific format for planners and policy-makers. This format is called the Census Transportation Planning Package (CTPP).

7.2 Introduction

The CTPP provides transportation planners and modelers with origin and destination (OD) data on commuters. This information can augment household travel surveys as an additional model development dataset for model estimation, calibration, or validation. In some cases, the CTPP is also used as a direct input dataset, as is the case with the Federal Transit Administration Simplified Trips-on-Project Software (STOPS) package.

The CTPP provides detailed information on commute trips between census geographies. It is available for state-to-state flows and traffic analysis zone (TAZ)-to-TAZ flows. The CTPP data also provide additional household characteristics of commuters like age, household structure, and income. These additional tabulations can be used to validate households by number of workers in models, and—in the case of the synthesized, or perturbed, data—used to validate worker flows by socioeconomic characteristics.

CTPP data development currently is funded by state departments of transportation through the American Association of State Highway and Transportation Officials (AASHTO). The data are most useful when combined with other datasets like household travel surveys or data from the Longitudinal Employer-Household Dynamics (LEHD) datasets offered by the U.S. Census Bureau.

LEHD offers commuter flows derived primarily from state unemployment insurance programs. The state unemployment insurance program has complete records on all private civilian jobs and workers. Since LEHD is based on a full accounting of jobs, it offers a more complete picture of OD coverage than the CTPP, which may miss low frequency OD pairs due to sampling. LEHD (discussed in Chapter 6.0), however, lacks socioeconomic characteristics of the travelers contained in the CTPP package.

Most often, CTPP data serve as an independent source for commuter flow validation. Transportation models are typically developed and estimated from household travel surveys and other similar types of surveys. In most cases, the entire survey dataset is used to build the model estimations, which precludes their use in model validation. The CTPP data are a well-used and independent alternative that can be used to compare whether transportation model commuter flows by market segment are well represented in the model. In some regions, auto ownership models and other auxiliary models are estimated or calibrated from the CTPP and related

American Community Survey (ACS) data. However, CTPP does not include nonwork travel, limiting its use in many travel estimation tasks.

In addition to transportation planning, the CTPP is also useful for economic and land-use planning purposes. The information can provide useful statistics on the characteristics of the labor force working in an area, including how far workers travel to jobs in the region.

7.3 Overview of Data Collection

AASHTO develops the CTPP tabulations from data collected by the U.S. Census Bureau. Between 1970 and 2000, the CTPP data were derived from the long form in the decennial census. Since the long form was discontinued in 2005, the CTPP has been built from rolling surveys in the ACS data releases. The data are derived largely from the employment and JTW questions on the survey.

Each year, the U.S. Census Bureau surveys approximately 2.5% of U.S. households in the ACS. Using this information, the U.S. Census Bureau releases 1-year, 3-year, and 5-year ACS profiles. For the 3-year and 5-year releases, data are released by blending surveys collected over the reporting period. This blending allows the U.S. Census Bureau to produce statistically reliable estimates of population characteristics at small geographic units. Using these survey results, AASHTO has produced two ACS-based CTPP releases, one using a 3-year ACS sample (2006–2008) and another using a 5-year ACS sample (2006–2010). The next anticipated release will use the 5-year ACS sample (2012–2016).

Based on the answers to questions about work location in the ACS, AASHTO can derive OD matrices down to small-area geographies. On the ACS form, respondents are prompted for the address of their work location. This information is aggregated by the U.S. Census Bureau and AASHTO to higher-level geographies for reporting purposes. This information is also cross-tabulated with mode of transportation and other socioeconomic information to build commuting market segments.

The CTPP is offered in two formats, "real" and perturbed. AASHTO works with the U.S. Census Bureau to ensure privacy is maintained for all geographic levels in the package. In the "real" data, privacy is maintained by suppressing data points that could lead to individual identification. As a result, fewer cross-tabulations are available in these data. In the perturbed dataset, AASHTO uses statistical techniques to infer reasonable proxies for suppressed values. Perturbing the data tables enables AASHTO to release useful cross-tabulations of commuter flows at small geographic levels.[9]

7.4 Procedures Used to Prepare Origin-Destination Tables

The CTPP website provides traveler flows by OD for download at several geographic levels. The commuter flow data can also be cross-tabulated with other socioeconomic characteristics before they are downloaded. In many instances, the CTPP data will be downloaded in a county-to-county or large subcounty zone (e.g., PUMA-to-PUMA, tract-to-tract) matrix to compare with travel model

[9] More information on data suppression and perturbing is available in NCHRP 8-79: *Producing Transportation Data Products form the American Community Survey that Comply with Disclosure Rules*.

results. Travel model OD tables can be aggregated to county-to-county and tract-to-tract flows for easy validation and comparison.

The goal of comparing CTPP to model results is to ensure that the model adequately reflects the commuting markets described by the CTPP data. The analyst, however, needs to pay special attention to the temporal dimension of the travel model and the comparable CTPP data. The CTPP data via the ACS are collected on a 5-year rolling basis. As such, “multiyear estimates describe the population and characteristics of an area for the full [5-year] period, not for any specific day, period, or year within the multiyear time period.” (U.S. Census Bureau 2008) Travel models often reflect a fixed point in time, like an average day in a particular year. As a result of the temporal differences, analysts may consider comparing shares of travelers or scaling the CTPP based on Census Population Estimates for consistency with the travel model’s time period. (Jeon et al. 2014)

7.5 Assessment of Data Suitability

7.5.1 Strengths

The CTPP offers a rich, readily available data source for planning agencies to validate models of work travel. The data from the 5-year ACS sample are available for geographies across the United States—although these data may be perturbed in some cases to protect privacy. These data offer a consistent approach to model calibration across the country, reducing the cost and time required to develop work trip OD trip tables.

7.5.2 Shortcomings

Several shortcomings exist with the CTPP data and require additional consideration when using the data for estimating or calibrating travel models. These shortcomings require context and expert opinion to guide their use in travel modeling.

Issues to be considered include the following:

1. CTPP data are useful for calibrating travel models for work purposes and understanding long-term historical trends, but the temporal data lag makes microtrend analysis or real-time assessments difficult. The U.S. Census Bureau does not recommend comparing overlapping ACS data series. This requirement limits CTPP releases to 5-year intervals. AASHTO uses the 5-year ACS data series to maximize sample size and provide data at the smallest possible geographies.
2. Because the CTPP represents surveys collected over a 5-year period, it is difficult to understand what the survey results represent. If comparisons are made to transportation model data, which represent a specific point-in-time estimate, then the CTPP data should be scaled to represent a similar point in time.
3. In addition to trip mode and purpose (i.e. work trip), the CTPP reports the locations of workers’ residences and workplaces. This constraint is inherited from the travel questions in the ACS and older long-form census survey instruments. This constraint limits the scope in which the CTPP data can be used for model estimation, calibration, or validation.

4. The CTPP includes information related to the LEHD worker flows and other similar data packages. Each package offers a windowed view of the underlying travel behavior based on the primary purpose of each data source. Often, each of these views can create conflicting assessments of the overall travel behavior in a location. Practitioners must understand the source of each dataset, its strengths, and identify the data most useful or relevant to the task at hand. The concerns also apply to ACS data used in the CTPP (e.g., labor force participation).
5. The margin of error in the CTPP is both a strength and weakness. As a strength, the CTPP transparently reports margin of error in its data products. As a weakness, however, particularly in small or sparsely populated geographies, the margin of error can be large, reducing the effectiveness of the data. Perturbing the data tables for generating cross-tabulations of commuter flows at small geographic levels also generates potential errors in the data.

7.5.3 Opportunities for Improvement

Big Data will transform the way more traditional surveys are collected and used in the future. Until now, the CTPP has relied on building estimates from representative samples of the U.S. population through the ACS. These survey formats are expensive to administer and preclude releasing precise data at small-area geographies due to sample sizes and privacy laws. On the other hand, Big Data relies on much larger samples of individuals through passive sensors (e.g., smartphones, mobile devices), but these larger samples may not be a representative sample of the population, which is one of the biggest challenges. As Big Data sources continue to evolve, AASHTO will have opportunities to reduce the cost or improve precision of the CTPP by augmenting census data with Big Data sources.

8.0 Visitor Survey Data

8.1 Background

Visitor travel surveys are a specialized type of travel survey focused on the tourism and visitor market in a region or study area. These types of surveys are used to develop a more complete representation of travel in many regions. In localized areas near tourist destinations like theme parks or major museums, the effect of visitor travel can be greater. For example, a recent study in Florida estimated that visitors to the state generated at least 10% of the vehicle miles traveled (VMT) in the state. In these situations, targeted visitor surveys inform planners and policy-makers about the travel behavior of visitors, and the results can guide infrastructure and policy decisions that may have been overlooked in a traditional household travel survey.

8.2 Introduction

Visitor surveys for transportation modeling purposes are conducted in a similar fashion to resident household travel behavior surveys. The surveys are collected either by traveler recall or in real time with a smartphone-based GPS travel survey app. Like other survey methods, including resident travel surveys and external or establishment surveys, the visitor travel survey is a method to quantify the scale and characteristics of visitor travel within a region or study area, including origin-destination (OD) characteristics.

Visitors are typically defined as people in the region or study area without a primary work or home location in the region or study area. Visitors fall into two major categories: business and leisure. This definition of visitors excludes long-distance or interregional commuters because they have a regular work location in the study area. Visitors and long-distance commuters are usually handled separately. The travel needs and economic concerns of visitors and long-distance commuters are often different, and regular commuters are sometimes segmented separately to focus on their unique political or economic concerns.

8.3 Overview of Data Collection

This section provides a general overview of the tools and techniques available to conduct a visitor travel survey.

8.3.1 Survey Universe Identification

Identifying the survey universe is a critical first step to completing a visitor travel survey. The universe is defined by the number of individuals in the overall area-of-interest population. In the case of a visitor survey, this is typically nonresidents traveling in the region. The visitor definition excludes interregional commuters who do not reside in the region, but travel into the region regularly for work. Long-distance commuters can be surveyed separately through an external travel survey.

Depending on the goals of the survey, visitor surveys can also be directed at residents within the region. For example, a recent Chicago Area Visitor Survey included residents living in the Chicago metropolitan region. This approach was chosen because the Chicago-area transit agencies wanted to tailor marketing, fare, and service strategies to better meet the needs of visitors to Chicago’s major tourist destinations regardless of the visitor’s home location.

8.3.2 Survey Questionnaire

The survey's primary purpose is to gain insight on the number and frequency of trips, location of trips, mode of trips, time period of trips (daily, weekly, monthly), and other trip characteristics along with information about the socioeconomic status of the respondent. Typical visitor surveys ask travelers to describe the characteristics of their trips and provide the following types of information.

- Trip Characteristics:
  - Origin and destination location of the trip.
  - Origin and destination purpose of the trip.
  - Travel mode(s) of the trip.
  - If transit is used, the access and egress mode and boarding and alighting station.
  - Time of day.
  - Duration of the trip.
  - Route, if possible.
  - Taxi, transportation network company (TNC) (e.g., Lyft, Uber), or transit fare.
  - Trip costs, including tolls, transit fares, TNC ride costs, etc.
  - Number of and relationship among participants.
- Socioeconomic Characteristics:
  - Income.
  - Ethnicity.
  - Age.
  - Household size.
- Visitor characteristics
  - Location and type of lodging (e.g., hotel\motel, residence, campground).
  - Length of stay of visit.
  - Purpose of visit (business, personal\recreational, or both).
  - Mode and location of entry into region.
  - Rental car availability.

Depending on the goals of the survey, it may also be useful to ask respondents about travel perceptions. This information can help planners and policy-makers better understand visitors' impediments to travel in a region. For example, visitors may choose TNC services over transit for personal reasons, but they may also choose TNC services because the transit system in a region is difficult for outsiders to understand. The traditional travel survey would reveal these behavioral choices, but it would be difficult to accurately deduce the justification for the behavior.

8.3.3 Recruitment Methods

8.3.3.1 *Airport Intercepts*

Administering a visitor survey in coordination with a local airport authority is one common way to reach regional tourists and visitors. In many cases, a survey is conducted at an airport gate while passengers are waiting to board their return flights. Visitors can be isolated in the survey by asking whether they live or work in the study area. The last full day of travel in the region is typically collected to minimize recall biases and capture a complete day of activities.

This approach is beneficial because it can be completed efficiently and waiting passengers are often a captive audience. The primary disadvantage of this type of survey is the significant potential to miss tourists or visitors who arrive in the region by car, train, or other nonair modes. Surveys also could be conducted at similar locations such as train stations or bus terminals. However, these additional surveys require more resources while only marginally improving the sample set. These additional surveys also ignore visitors entering the study area via private automobiles.

8.3.3.2 *Lodging and Tourist Destination Intercepts*

Visitor intercepts can also be conducted at key locations where tourists and visitors congregate. These locations include hotels, tourist destinations, or large convention centers. The primary advantage to this approach is that it avoids travel mode bias such as in the airport intercept surveys. Visitor intercept surveys do have three major disadvantages. First, depending on the number of sites selected and the geographic distribution, the costs associated with recruiting participants could be quite high. Second, visitor engagement is a concern, since visitors do not have much interest in participating in a travel survey when they are on their way to tourist destinations or in the process of checking into or out of a hotel. Third, it may be difficult to properly weight or expand the resulting data since the chosen survey locations may not be fully representative of the full set of tourist destinations.

8.3.3.3 *Passive Recruitment (Advertising)*

To increase response rates and lower the burden associated with answering a survey immediately, visitor survey intercepts can be enhanced by passively recruiting people in a targeted advertising campaign. Advertising can direct visitors to an online survey to detail the previous day's travel. This helps overcome cases where recruiting in hotels and other visitor-oriented locations is difficult or impossible. It also allows visitors to take the survey on their own time. For example, in 2011, Oregon Metro conducted a visitor survey and asked 30 hotels in downtown Portland to assist with a visitor intercept survey. Only nine hotels agreed to let surveyors intercept their customers. In these cases, recruitment can be done through targeted advertising. This could involve signage at visitor-oriented destinations or targeted online advertising inviting participants to sign up on a website. In the Portland example, Metro distributed postcards at the airport, convention center, and transit facilities across the region and used online advertising to increase the response rates.

8.3.4 Incentives

Incentives are a common part of travel surveys at the national, state, and local levels. Research has demonstrated that incentives are an effective way to boost response rates. Visitor surveys suffer from lower response rates than resident surveys because most visitors are on vacation or on their way to business meetings. Often, visitors in these situations do not want to participate in a long survey. Visitors are also unlikely to have a personal stake in the long-term planning of a region, so the perceived return on time investment may be quite low.

Incentives such as gift cards are one way to increase the response rate in visitor surveys. Incentives can be distributed to each individual or household that completes a valid, verified survey. The individual gift cards typically range in value from $10 to $40 for each completed survey. Some surveys use a grand prize incentive where respondents are entered into a larger grand prize drawing after completing a valid survey. The grand prize is often of more substantial value and may be oriented toward visitors; prizes may include airline tickets, hotel vouchers, or complimentary theme park tickets. In some cases, surveys offer both individual incentives and a larger grand prize lottery. Incentives do increase the cost of surveys; however, incentives are often more cost-effective than increasing the entire recruitment pool to boost the number of valid and complete surveys.

8.3.5 Seasonality

The timing and location of the recruitment should account for the seasonality of visitors to the region. For example, in the Phoenix region, the visitor population is significantly impacted by the arrival of “snowbirds” in the winter months, the commencement of Major League Baseball’s Spring Training program in the spring, and the desert heat in the summer months. Conducting a visitor survey in January, March, and July in Phoenix will yield different perspectives of the visitor travel market.

Recruitment locations could also change depending on the time of year. In some situations, conducting the survey on a rolling basis throughout the year to capture different visitor markets might make sense. If the survey is conducted over a prolonged period, then the scaling or weighting of the survey results should consider scaling factors appropriate to the season. Specifically, the scaling should reflect the overall market throughout the year, and the scaling should account for variation in the market segments throughout the year. For example, summer months may be dominated by recreational tourists while other parts of the year may be more oriented toward conventioneers and business visitors.

8.3.6 Collection Methods

8.3.6.1 *Smartphone-based GPS Travel Survey Application*

Smartphone-based GPS travel survey methods are the most accurate method for collecting travel survey data. Once respondents are recruited, they are prompted to download an app from the appropriate store for their mobile device’s operating system. The app, once enabled by the user, will passively track movements, and it will prompt the user for travel details once it senses the trip is completed. This platform provides the cleanest and most complete data. Because the data are collected in real time and instantly validated online, inconsistencies can be flagged and resolved on the spot with the respondent. This method also prompts the users to enter information about smaller trips that may have been forgotten or underreported in other methods. This reduces the recall bias for this method more than other traditional approaches.

8.3.6.2 *Computer-Assisted Interview or Online Recall*

This method uses an interview to assist survey respondents in recalling the details of their travel. The interview can be done in person or with the interviewer via phone. In some instances, the interview is conducted completely online using a web-based survey without the assistance of a human interviewer. The interview often covers travel that occurred on the current day or the previous day. Recall bias becomes more significant for each preceding day the survey attempts to recall. This survey uses a computer program to prompt questions for the respondents. Like smartphone survey data collection, the respondent's answers are checked in real time, which allows the surveyor to follow up with additional questions to address errors or inconsistencies. This method, like the smartphone-based GPS travel survey method, also allows for the survey to branch and lead toward more probing questions (e.g., fare questions for transit trips but not walking trips), when warranted.

8.3.6.3 *Paper-based Recall*

The paper-based recall method uses a travel survey printed on a piece of paper. Respondents are recruited to complete the survey on their own or with the assistance of an interviewer. The respondents are asked questions about their travel behavior on the current day or the previous day. The respondent's answers are recorded on a piece of paper and later entered into a computer database. These forms can be distributed in large quantities by a small number of staff, reducing the expense of fielding a survey. This type of survey collection is the cheapest and easiest to administer, but many surveys will be rejected during the data validation phase. While being subject to the same recall bias of computer-based recall surveys, the paper-based recall offers few opportunities to intervene with the respondent if their responses include inconsistencies or errors.

8.4 Procedures Used to Prepare Origin-Destination Tables

8.4.1 Data Validation

Before using the survey data to develop an OD database, the completed surveys should be reviewed for accuracy and completeness. The inaccurate and incomplete survey results either must be corrected or discarded from the final database. A smartphone-based GPS travel survey offers the greatest opportunity to capture errors in real time and prompt users for correction. Pencil and paper recall surveys will be the most prone to poor recall and require the most data cleaning.

The following example data checks are illustrative of what could be run to verify and clean the data:

- Validate Locations and Addresses: Ensure trips have origins and destinations at valid locations.
- Route and Origin-Destination Consistency: Verify the route reported is a valid route for the specified mode between the origin and destination.

- Purpose and Location Consistency: Verify that the trip purpose and the destination are consistent. For example, ensure that dining trips have a destination at a restaurant.
- Travel Validation: Ensure that the travel time reported between origin and destination is reasonable based on the mode, route, and time of day.
- Dwell Time: Verify that time spent at a location is reasonable based on the trip purpose and type of location.
- Trip Chaining Consistency: Verify that the reported departure time is after the arrival time of the previous trip destination. This also includes verifying the destination of the previous trip is consistent with the origin of the current trip.
- Mode Consistency: Verify that modes used throughout the trip are logically consistent. If a person walked to the bus for their outbound trip, that person most likely does not have a bike or car available for the next trip.
- Time-of-Day Validation: Ensure that trips are occurring at logical times in the survey. For example, it is unlikely for an elementary school trip type to leave home late at night.

8.5 Data Use in Travel Forecasting Models

The data collected in the visitor survey are used to supplement the existing traveler survey data and to estimate and calibrate total travel in the region. Because any visitor model or special generator model works in coordination with other travel model components, the calibration should be done in parallel with other models in the travel modeling chain, including trip generation, trip distribution, and destination choice models.

8.5.1 Scaling

The final, cleaned survey data is useful on its own, as an input to trip generation or trip distribution. In the context of travel model calibration, visitor survey data should be expanded or weighted to represent the entire population of visitors in the study area to make the survey data useful for transportation modeling. For visitor surveys, this may be the most difficult portion of the survey process because accurate measurements of the visitor population may not be readily available, especially detailed demographic statistics of travelers. For a regional visitor survey of travel, a diverse sampling plan including geographic, facility type, and customer type variation will be important to minimize sampling biases around socioeconomic characteristics like race/ethnicity, gender, age, and income. Estimates of the total number of visitors might be kept by a local tourism authority or other regional marketing organization. If the survey recruitment was conducted primarily at tourist destinations, then the scaling factors could be derived from the average daily attendance at that location. Another potential source of scaling is passively collected, Big Data sources. Some transportation data vendors produce estimates of visitors in a region by analyzing observed, usual trip-making patterns using personal mobile devices or factory-installed, onboard GPS devices. The new sources of data will continue to improve as collection biases are actively addressed by the vendors. Finally, the scaling factors should consider the seasonality of visitor travel patterns in the study area.

8.5.2 Activity-Based Visitor Travel Model

If the survey data are sufficiently detailed, then the visitor survey data can be used to build and estimate a disaggregate, visitor travel model for both conventional, 4-step models and activity-based models. The survey can be used to build a disaggregate population of visitors with detailed socioeconomic characteristics like residents in a traditional resident based AB model. The visitor survey can then be used to estimate visitor travel behavior including visitor-specific travel characteristics (e.g., trip purposes, rental car use) and propensity to travel to specific tourism destinations across the region. Based on the survey results, special attractor size terms can be added to the destination choice models to generate an appropriate number of trips to visitor facilities. Most of the remaining model components (e.g., time-of-day choice, mode choice) are derived consistent with model estimation of a resident activity model.

In 2011, the San Diego Association of Governments (SANDAG) used an intercept survey of airport passengers and hotel guests to build an AB model of visitor travel to complement the resident AB travel model. In the San Diego visitor model, the SANDAG survey revealed that approximately one out of three visitors had a personal vehicle available (e.g., personal car or rental car). Because most visitors did not have a car available, walk and taxi mode shares in the survey results were far higher than the resident travel survey. SANDAG added an auto availability attribute onto the simulated visitor records. This flag was then used by the tour and trip mode choices models as a choice constraint for travelers. For this reason, the SANDAG visitor model also added a taxi mode to the visitor model mode choice structure.

8.5.3 Other Policy Analytics

Finally, visitor survey data has numerous uses outside of transportation planning analysis. The information could be used by local planners and economic development officials to better understand trip locations, trip characteristics, and traveler characteristics for future development opportunities, marketing purposes, and traffic analyses near existing tourist destinations.

8.6 Assessment of Data Suitability

8.6.1 Strengths

As discussed, tourism generates significant travel in many regions, and this source of travel demand is important to capture in travel demand models. These visitors use the transit and transportation infrastructure of the city and surrounding region. (Bellafonte 2017) This nonresident travel is not captured by traditional household surveys, and visitor surveys have been developed to address this gap.

Visitor travel survey data can be used to build traffic to and from special generators (e.g., airports, hotels, convention centers, theme parks). In the simplest form, the information in the survey can be scaled to the volumes at the survey locations and OD tables can be generated by time of day, mode, and purpose. Depending on the application, these trip tables could be scaled with population or employment growth in the region for future-year conditions.

In a more complex analysis, the visitor survey data could be used to estimate a separate list of production and attractions for trips to and from the special generator sites. This list of production and attractions would then be taken through the traditional estimation of trip distribution and mode choice. The final trip tables from mode choice would be fed into assignment as either their own matrix or added to existing matrices from the resident model. After assignment, traffic counts and transit boarding could be evaluated.

8.6.2 Shortcomings

Shortcomings with visitor survey data require additional consideration when using the data for estimating travel models or other planning analyses. These shortcomings require context and expert opinion to guide travel modeling applications. The following issues require consideration:

1. Visitor surveys can be challenging to administer because the respondents are often not personally invested in the region or study area. This creates two primary problems: recall bias and survey fatigue. Because visitors are in the region or study area infrequently, the survey respondents may not be familiar with the geography of the region making it difficult to recall origins, destinations, and other trip characteristics in an interview. Visitors are also unlikely to have a personal stake in the long-term planning of a region, so visitor respondents are also more likely to have higher rates of survey fatigue. As a result, tools such as incentives need to be used to increase response rates.
2. Survey weighting plans are challenging to develop because the tourism and visitor market is difficult to define and identify. Survey controls are often only available at an aggregate level from tourism authorities, and the controls can be imprecise because visitors arriving by private vehicles are difficult to identify and survey outside of primary tourism destinations. Once the survey data are collected, care must be taken to validate and expand the survey results.
3. Visitor surveys tend to be biased toward travelers in central business districts or primary tourism destinations. The traditional intercept methods at tourist sites or major hotels may miss the opportunity to understand more budget-oriented, business travelers who may stay at hotel and motel locations closer to suburban office parks far from the urban core. The sampling plan must be based on the needs and uses of the survey.

8.6.3 Opportunities for Improvement

The introduction of new passively collected, Big Data sources over the last decade offers new opportunities to refine and improve visitor survey implementation. These additional data sources from mobile carriers, navigation devices, and other location services can help more effectively target the surveys to visitors. These sources can also be combined with traditional survey methods to provide a richer view of visitor travel within regions. Because passive data sources can allow analysts to infer a home location based on regular travel patterns, these data sources can also identify when individuals are traveling and visit other regions. These sources will continue to evolve, further segmenting travelers by more discrete purposes and socioeconomic characteristics. Big Data sources also offer an opportunity to better target intercepts of potential respondents in the field. Surveyors already use census data to target oversamples of travel modes or socioeconomic characteristics. Surveyors also use American Community Survey data to target neighborhoods for oversampling. Similarly, passive, Big Data sources could be used to better target intercept locations based on the observed travel patterns of visitors to the region over time.

9.0 On-Board Transit Survey Data

9.1 Background

A specialized type of origin-destination (OD) travel dataset may contain information on a single mode of travel rather than all person trips. This type of information can help practitioners understand the current market for that mode and forecast its future use. One notable example of mode-specific travel data is a transit OD dataset derived from a transit rider survey. This chapter describes transit rider surveys, the processing that is performed to convert this information into an OD dataset, and the resulting uses of this information.

9.2 Introduction

Transit rider surveys have been conducted for many years to gather information on the characteristics of transit users and their trip-making patterns. These data have been used to support several transit planning and management activities, ranging from assessments of customer satisfaction to calibrating transit demand forecasting models. Since transit rider surveys are frequently conducted while customers are traveling on buses or trains, these surveys are often called “on-board transit surveys.” However, similar information can be gathered by interviewing passengers at transit stations or bus stops or by recruiting survey-takers online.

Useful OD data come from a subset of transit rider surveys that are specifically designed to capture information on origin and destination location, other trip characteristics, and traveler characteristics. These surveys also require carefully collected information on total transit trip-making (often obtained by conducting an “on-to-off” count [described below], automatic passenger counters, or fare collection data); this information is the basis for expanding sample data to represent the total population. These surveys are sometimes called transit OD surveys to distinguish them from other transit survey efforts that are intended to collect more general information on customer satisfaction.

Because a survey intended for developing OD information requires *complete* information on both traveler/trip characteristics and on total trip-making, many survey efforts are separated into two key elements:

- **Transit OD Survey**. This element of the survey effort involves a paper-based or computerized survey questionnaire in which travelers are asked to describe in detail the trips they are making and to provide additional information about their own personal characteristics (e.g., sociodemographics).
- **Transit Control Data Gathering**. This element of the survey effort involves collecting information on the characteristics of the total population of transit travelers. It is used to weight (sometimes also called “expand,” as the weight corrects for oversampling or undersampling of a population and then expands the survey record up to the total number of trips it represents on the system) the results of the OD survey to represent total transit travel. The best form of survey control is a count of travelers for each combination of boarding and alighting station/stop pairs for each transit route. This information can come from fare gate information or from a specialized on-to-off count. If a robust sample or full on-to-off data cannot be collected, then less-robust information—such as trip boardings

and alightings by route segment, time of day, and direction—may be sufficient for use as a basis for OD weighting.

9.3 Overview of Data Collection

This section provides a brief overview of how transit OD data are collected.

9.3.1 Survey Questionnaire

Typical transit OD surveys ask passengers several questions. Passengers are asked to describe the characteristics of their trips and to provide information on their individual and household characteristics. To generate a useful database of OD travel, these questionnaires must gather information on the following characteristics:

- Trip characteristics:
  - Origin and destination locations at a level-of-detail sufficient to determine the origin and destination traffic analysis zones (TAZs). If possible, origins and destinations should be geocoded to latitude and longitude (to within 1/10,000 of a degree) to allow assignment to alternate systems of TAZs that may be defined at some later date.
  - Access mode used to travel from the trip origin to the first boarding location.
  - Egress mode used to travel from the last alighting location to the trip destination.
  - Boarding and alighting stop identification and route identification for the current unlinked trip (i.e., the specific bus or train segment where the passenger is surveyed). If possible, boarding stop, alighting stop, and route identification for all trip segments should be collected.
  - Number of unlinked trips that comprise the current origin-to-destination linked trip.
  - Purpose of the trip and identification of whether the trip origin or destination is home, the workplace, school, shopping, or other kind of place.
- Traveler/household characteristics:
  - Household income classification.
  - Number of operable vehicles owned by the household.
  - Number of working adults in the household.

These questionnaires can be administered in several ways. In some cases, two or more of these survey administration techniques can be combined to provide a convenient array of options for travelers and increase overall response rates.

The most recent successes have been achieved by having trained interviewers ask travelers orally about trip and household characteristics and record the answers immediately in a tablet-based data entry form. Interviewers help improve the quality of the data collection since they can help explain the questionnaire if respondents are confused and can help clarify responses that are clearly erroneous. They can also assist respondents with the tablet-based procedures that identify trip origin and destination locations—often, the most difficult-to-answer questions encountered on these surveys. Finally, one of the biggest benefits of in-person interviews is that a response rate (the percentage of randomly approached people who provide complete information) greater than 70% is often achieved. Experience has shown that few riders decline to take a survey when asked face-to-face by an interviewer who has been properly trained, which reduces—although does not eliminate—the effect of a nonresponse bias. This is important, as other survey methods tend to have much higher nonresponse rates, which can raise concerns that survey results are not representative of the characteristics of the full population of transit riders.

Although interview-based questionnaires are generally the most accurate, concerns may include higher costs or the lack of consistency with past, successful surveys. To address such concerns, surveys may be collected with traditional paper-based surveys distributed in large quantities by a small number of staff, reducing the expense of fielding a survey. However, this approach can also result in lower response rates (causing higher nonresponse bias) and higher levels of incomplete or inaccurate responses. Unlike answers entered in real time into a tablet, paper-based responses are not checked in real time. As a result, many more survey responses will be judged to be unusable and eliminated from the final survey database.

Another approach for collecting rider surveys involves e-mailing riders a URL for an online survey questionnaire and asking each traveler to log on to a survey website and answer questions about their trip later. This approach offers real-time logic-checking but also suffers from lower response rates and possible memory lapses regarding the trip being recorded. This technique can also result in a skewed response (e.g., higher-income riders might respond in greater proportions). It also requires that the transit agency have many e-mail addresses for its ridership base.

An important consideration for any of these survey techniques concerns the number of records required to generate a useful dataset. Experience in this field suggests that a 10% sample is sufficient in most cases to generate a useful sample. This guidance is not based on statistical tests. This guidance instead comes from experience that a 10% sample is sufficient to develop an OD transit trip table at the TAZ level-of-detail that is helpful to support travel forecasting model development and application. One problem with a simple 10% rule is that it might involve collecting more information than is needed for uninteresting situations while collecting insufficient data for more relevant travel.

9.3.2 Control Data and Weighting

All data collection efforts that attempt to use a sample to represent an entire population must also develop a means for scaling the sample information up to represent all users. For transit rider surveys, this is usually accomplished by attaching a weight to each record that indicates how many trips in the population are represented by that record. Separate weights are usually defined to scale the record to match unlinked trips (i.e., each boarding) and linked trips (i.e., the overall journey from the trip origin to the trip destination). These weights must account for the fraction of all passengers that successfully respond to a survey are not proportionately distributed over the entire system. For instance, travelers making short trips on crowded transit vehicles have a lower chance of being approached by a surveyor and—when they are—they may not have adequate time to complete the survey before they alight from the transit vehicle. Some groups of travelers may be more or less likely to respond to surveys depending on their trust of governmental institutions, level of education, and other factors. To the extent possible, survey weighting should correct for these response biases so that the weighted surveys represent the full traveling population as accurately as possible.

One way to improve the survey representation of the transit population is to weight the survey to match the observed number of trips for each on-to-off pair. This count is more than just how many riders get on or off at a given stop (Figure 8). It also encompasses how many riders make a trip from stop A to stop B on a transit vehicle as it proceeds along its route (Figure 9). The representation in Figure 9 is the data that are used for survey expansion.

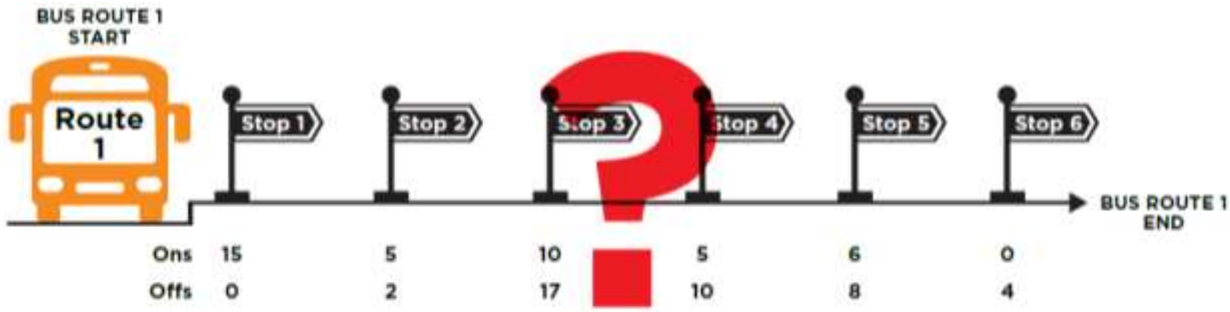


Figure 8. APC data provide on-to-off counts but *not* boarding and alighting pair data.

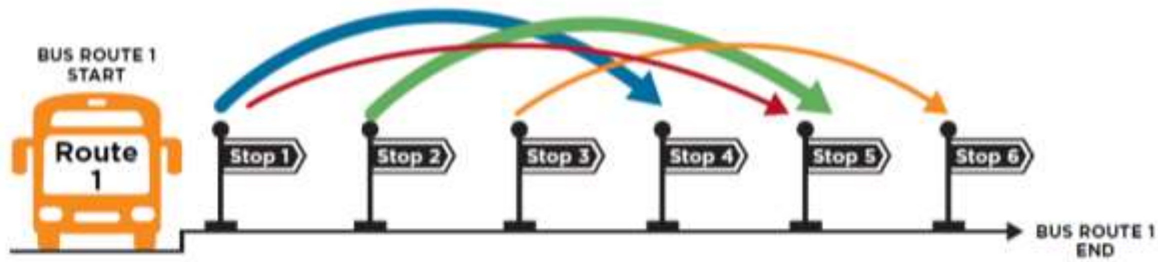


Figure 9. An on-to-off count provides the boarding and alighting pair data (or flows).

This information is most commonly obtained from a separate on-to-off survey.[10] This type of survey is typically conducted for approximately 20% of all scheduled transit vehicle trips. On each sampled vehicle trip, a short survey card is distributed to all boarding passengers and then collected when each passenger disembarks. The survey card may have just two questions (where did you board and where did you alight?). Alternately, the survey may be barcoded and scanned by members of the survey crew at the time of boarding and at the time of alighting. These times can be converted to latitudes and longitudes using information on bus/train location from an Automatic Vehicle Location system. Because the survey is easy to complete, on-to-off count response rates are usually quite high, often 90% or more.

These boarding and alighting pairs for sampled trips are then weighted up to the total number of riders at each stop on each route using count information obtained from automated passenger counting (APC) systems by weekday time of day based on an average of APC data over multiple

[10] If a system uses fare cards to control system entrances and exits, then fare data may also be used to estimate on-to-off ridership.

weekdays. Factoring is typically accomplished using an iterative proportional fitting algorithm. With this method, on-to-off count results are used as an initial "seed" matrix. Stop-level APC data provide the overall row and column control totals. Rows and columns are then iteratively factored to match row and column totals until both rows and columns match the desired control totals. Figure 10, below, illustrates this process.

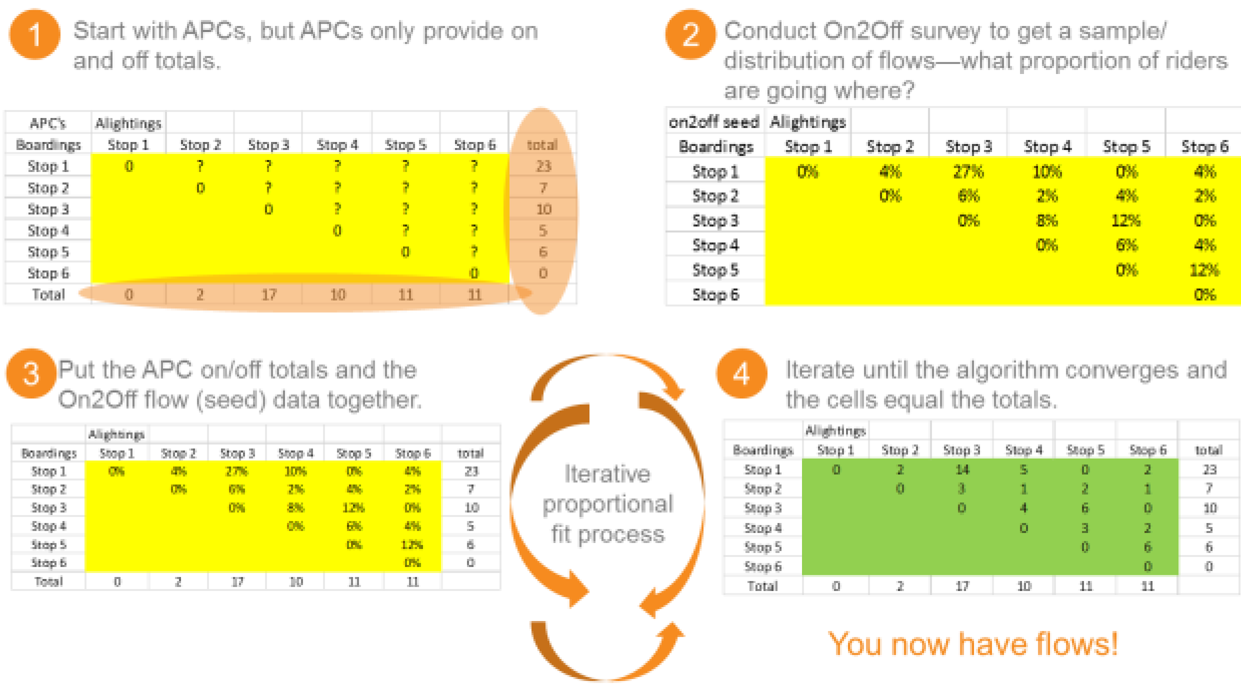


Figure 10. Visual depiction of survey weighting process.

If an on-to-off count is not feasible, then a fallback position for survey control is to obtain counts by route, direction, time of day, and (if possible) boarding/alighting counts by stop and route.

9.4 Procedures Used to Prepare Origin-Destination Tables

Field work is followed by a series of steps to review data quality and convert raw survey responses into a useful database of transit origins and destinations. These steps are described in the following subsections.

9.4.1 Data Validation

Before being used to develop an OD database, survey responses must be reviewed for accuracy and completeness and unusable records must be culled from the dataset. Data validation checks include the following:

- Confirm valid response for key fields, including origin location, destination location, boarding stop, alighting stop, route(s) used, trip purpose, home location, and socioeconomic characteristics.
- Check consistency of OD location with bus routes and access/egress modes.
- Check against incorrect round-trip reporting.

9.4.2 Survey Weighting

The exact form of the survey weighting to expand questionnaire responses (a sample) to represent the entire population depends on both the sampling plan and the availability of control data. The basic process involves weighting in two stages:

1. Expand survey results to represent unlinked trips on the vehicle where the survey was conducted. The exact nature of this calculation depends on whether on-to-off counts are available or whether data are only available at the route level (stratified by direction and time of day) or boarding or alighting location. In any case, the expansion factor (weight) is equal to counted trips for the expansion frame divided by the number of survey responses. Similar procedures are employed if the survey is conducted on a station platform except that the control is by station boarding rather than route.[11]
2. Convert unlinked trip weights to linked trip weights. The unlinked trip weights are next converted to linked trips by dividing by the number of boardings (number of transfers plus one) made during the entire linked trip. The linked trip weight is used for development of the OD dataset.

9.4.3 Supplemental Processing

Additional survey processing may be required prior to development of an OD database. One of the first steps is dependent on whether the table should be organized in terms of origins and destinations (OD table) or whether the table should be organized as productions and attractions (PA table). In this context, the production location is the origin or destination location that corresponds to the traveler's home. The attraction location is the other end of the trip. In the case of non-home-based trips, productions are assumed to equal origins and attractions are assumed to equal destinations. As an illustration of the meaning of OD and PA tables, imagine a worker traveling from home in the suburbs to a job in the city and then returning home in the evening. This example trip would produce the following records:

- **OD Table Structure**:
    - One work trip from the suburban location to the city location.
    - One work trip from the city location to the suburban location.
- **PA Table Structure**:
    - Two work trips from the suburban location to the city location.

Most transit analysis is conducted with PA tables since it keeps the home end of the trip separate from the nonhome end. This concept allows models to properly associate the surveyed trip with characteristics associated with the residential location (e.g., household income, availability of a car for access) and the work location (e.g., employment type).

[11] Occasionally, trips exist in the control data for a given boarding/alighting pair but do not exist in the OD survey dataset. When this happens, adjacent stops should be combined into groups of stops and the survey expansion should be conducted for stop-to-stop groups rather than for individual stop-to-stop pairs.

Additional survey processing is required if a PA table is desired. This involves determining whether the destination end of the trip is the traveler's home. If it is, then all the characteristics associated with the destination (e.g., destination location, destination type, and egress mode) are stored in the production fields and all the characteristics of the trip origin (origin location, origin type, access mode) are stored in the attraction fields. If not, then the reverse is true—all the characteristics associated with the origin are stored in the production fields, and all the characteristics of the trip destination are stored in the attraction fields.

Supplemental processing may also be required to classify mode of access, trip purpose, socioeconomic characteristics, and path according to the desired (often model-specified) definitions. This step involves converting whatever fields were defined in the survey to the definitions defined for modeling or analysis purposes. Typical fields are defined as follows:

- Mode of access (production end):
    a. Walk.
    b. Kiss-and-ride.
    c. Park-and-ride.
- Trip purpose:
    a. Home-based work.
    b. Home-based other (may be further subdivided into shop, school, social-recreation, and other).
    c. Non-home-based.
- Socioeconomic classification:
    a. 0, 1, or 2 or more household cars.
    b. Household income group.
    c. Measure of automobile sufficiency.
- Transit path:
    a. Commuter rail.
    b. Urban rail.
    c. Bus.

9.4.4 Preparation of Origin-Destination Tables

OD databases are prepared by accumulating linked transit trip weights across all survey records for each combination of production zone, attraction zone, trip purpose, socioeconomic class, access mode, and transit path.

9.5 Data Use in Travel Forecasting Models

A comprehensive transit database can be used in three types of analysis once it is assembled. The following subsections discuss these applications.

9.5.1 Development/Calibration of Conventional Travel Demand Models

One key use of transit OD information is to help develop and calibrate conventional travel demand forecasting models.

For example, an assignment of transit survey data to transit networks can be used to simultaneously check the validity of the OD data and the networks that are used to represent transit supply. This is accomplished by converting transit OD data (in PA table format) into travel model-type trip tables. These trip tables are then assigned to the transit network using the transit assignment procedures of the local model. These assignment procedures generate a listing of modeled (from the survey, in this case) station boardings, alightings, and route ridership that can be compared to independent count data. If the assigned survey volumes match ridership counts well, then the analyst can have confidence in the survey, the transit networks, and the path processing procedures. If assignment results do not closely match counts, then each step must be reviewed to determine what actions must be taken to improve the representation of how travelers utilize the existing supply.

The survey OD tables can be used to assess the validity of the demand elements of the model after the transit trip table is successfully assigned to transit networks and the accuracy of the trip tables and transit supply are established. This is done by comparing the transit person trips generated by the mode split model to the survey trips. This comparison should be performed separately for each production district (aggregation of TAZs), attraction district, socioeconomic class (auto ownership, income, or auto sufficiency), trip purpose, access mode, and transit submode. If mismatches are found between the mode split model output and the survey data, then the analyst must examine both the mode choice model and key precursor models, including trip generation and distribution. This process often involves proposing a theory for the root cause of the mismatch, testing a solution, and comparing updated transit model results to the survey. Given the unknown nature of many problems, this investigation cycle may need to be repeated multiple times before an acceptable solution is found.

9.5.2 Direct Application in Incremental Models

Some types of transit models can use transit OD tables directly as inputs. These models are known as incremental models because they focus on representing the change (“increment”) in demand that results from a change to the transit system. One recent example of this type of model is the Federal Transit Administration Simplified Trips-on-Project Software (STOPS). Incremental models work by using the transit OD survey as the basis for all computations. If the existing transit service is the same as the service that operated at the time of the survey, then the transit trip table is equal to the result of the survey effort. If that table and the underlying transit networks and path processing procedures generate a realistic portrait of transit boardings, alightings, and route ridership, then the model is calibrated. When changes are made to the transit supply, zone-to-zone transit travel times will change, and the incremental model will predict a proportional change to the number of trips made by transit. In many cases, incremental models are much quicker to calibrate than conventional models. That is because incremental models skip the process of calibrating the trip generation, trip distribution, and mode choice model components, which are often the hardest steps associated with conventional transit model calibration. Although testing of the survey trip table, transit networks, and path procedures must still be done (and possibly revised), these elements are often more straightforward than calibration of the demand modeling portions of travel models.

9.5.3 Other Planning Analyses

Transit survey data have considerable utility for nonmodel planning analyses. Quantitative transit analysts often use information on trip locations, trip characteristics, and traveler characteristics as part of service planning, mobility, and social equity analyses.

9.6 Assessment of Data Suitability

This section discusses the strengths and shortcomings of transit rider survey information and presents opportunities to improve this data source.

9.6.1 Strengths

The fundamental strength of transit rider survey data and the resulting OD flows is that these data represent transit trip-making patterns using characteristics of each trip as reported by the travelers. Expansion procedures can minimize nonresponse bias. The process of converting these data to assignable trip tables and then comparing assignment results to counts confirms the accuracy of all steps in the process—survey data processing, transit network coding, and path-building. When these review steps are successful, users can have high confidence in the accuracy of the OD database.

9.6.2 Shortcomings

The most significant shortcoming of this data source is that a complete survey of a large transit operator can take over one year to accomplish and require considerable financial resources. To date, the most useful surveys have collected system-wide samples of 10% of their customers. For large agencies, this could result in more than 50,000 survey responses and cost over one million dollars, if both an on-board OD survey and an on-to-off count are conducted. The high level of expense results in some agencies skipping controlled OD surveys altogether or foregoing certain parts of the process (e.g., not conducting an on-to-off count) while other agencies elect to collect this information rarely. As a result, these survey datasets are not available in all metropolitan areas. When they are available, some are too old or too inaccurate to be useful.

9.6.3 Opportunities for Improvement

The best opportunity to improve this data source is to reduce the cost of data collection to a point where it can be collected more frequently or even as a routine element of ongoing agency operations. The most promising option might be to merge survey-type data collection with data from each agency’s fare collection process. Possibilities exist for fare collection procedures (particularly online purchases) to collect limited information on utilization that can be used to update and reweight survey data in the interval between survey field data collection periods. As with other traditional data sources, the value of transit survey data can be extended by combining these data with other data sources. For example, mode choice model estimation can sometimes be improved when transit surveys are combined with household trip-diary surveys.

10.0 Bluetooth Data

10.1 Background

This chapter describes how data obtained through detecting Bluetooth devices can be used to determine origin-destination (OD) estimates and the resulting uses of this information. Additional background information on using Bluetooth data for OD analysis can be found in the FHWA study *Synopsis of New Methods and Technologies to Collect Origin-Destination (OD) Data*. (Hard et al. 2016).

10.2 Introduction

Bluetooth technology is a wireless communications system used in mobile phones, computers, personal digital assistants, cars, and other short-range wireless communications devices. Bluetooth technology operates by proximity—Bluetooth-enabled devices that are close to one another can connect to allow transmission of voice or data. For a connection to occur, each device needs to be in "discoverable" mode with Bluetooth enabled.

Bluetooth devices are rated as Type I (100-meter detection zone); Type II (10-meter detection zone); or Type III (1-meter detection zone) and can detect any other Bluetooth devices within their range. Type I detectors are appropriate for estimating OD information. All Bluetooth-enabled devices operate within a globally-available frequency band of 2.45 GHz.

Bluetooth devices emit a unique, 48-bit electronic identifier known as a Media Access Control (MAC) address, or MAC ID. The MAC ID is generated in two parts: the first half of the MAC ID is assigned to the device manufacturer, while the second half of the MAC ID is assigned to the specific device. While the MAC ID is unique to each Bluetooth device, it is not linked to an individual person.

Bluetooth detectors can be placed adjacent to highways to record the MAC IDs of devices that come within their detection range, along with a timestamp and date of the detection. The latitude and longitude of the detector are also recorded, which enables the device to be precisely located for measuring distance and travel time between deployed Bluetooth detectors.

Deploying Bluetooth detectors can be for permanent or temporary installations. Permanent installations must have electricity provided to the unit. Portable units are battery-powered devices that can be attached to utility or signposts near roadways. Battery-powered units with solar recharging capabilities are also available.

Bluetooth detectors can be outfitted with a wireless modem for transmitting the data to a central server. Some detectors store the data locally on a hard drive, which requires downloading after the deployment. Hybrid solutions are also used, where the data for each day are saved locally, then transmitted to a central server once each day.

As vehicles pass by, detectors wirelessly acquire the unique ID along with the date and time of the observation. By simultaneously deploying multiple detectors throughout a study area, vehicles can be tracked over time as they are detected at multiple detectors in sequence. The raw data provide the basis for estimating travel speeds and OD information.

10.3 Overview of Data Collection

Bluetooth detectors can be installed permanently or temporarily. Temporary deployments are usually for a minimum of three days. Longer deployments (for 1–2 weeks) are desirable since more data are collected. However, battery life of the units must be considered as well. Figure 11 consists of two photographs of a Bluetooth detector, showing an open case (left) and a case installed on a sign adjacent to a highway (right).

Source: RSG.

Figure 11. Photograph of a Bluetooth detector box (left) and a typical installation (right).

The percentage of vehicles with discoverable Bluetooth devices varies, but in recent studies has ranged from 4-15%. For this reason, **traffic counts must be obtained simultaneously with the Bluetooth deployment to determine the expansion factor necessary to represent the full population of vehicles**. Figure 12 shows a layout of Bluetooth detectors for an OD study conducted along I-95 for the Florida Turnpike Enterprise. For this study, the layout plan was devised to estimate interchange-to-interchange travel along the corridor. A total of 32 Bluetooth detectors were deployed.

Source: RSG, using ArcMap.

Figure 12. Map showing temporary Bluetooth detector locations along I-95 in Florida.

10.4 Procedures Used to Prepare Origin-Destination Tables

After Bluetooth data collection, the raw data must be processed to construct vehicle trips. A common approach is to construct vehicle trajectories by grouping detections together according to the Bluetooth device ID of each detection, then to sort each group according to the date and time. The resulting trajectories each correspond to observations of vehicle trips over the course of the data collection period.

The trajectories should be further processed for two reasons. First, detectors may record multiple detections from the same device, which happens when a vehicle is within range of a detector for more than a few seconds. This situation could occur, for example, when a vehicle is stopped at an intersection. To eliminate duplicates, groups of contiguous detections that all occurred at the same detector within some time threshold (e.g., three minutes) of each other can be combined into one representative detection (clustering). Good practice is to use the first instance of detection in this chain, as it represents the earliest arrival time, and to eliminate the remaining detections from the analysis.

Another issue to investigate during data processing is that some trajectories represent more than one trip, which can happen when the same vehicle makes multiple trips within a study area during the deployment period. Again, travel time thresholds (e.g., five minutes) can be used to split trajectories into component trips. These thresholds specify the maximum time that could elapse between two adjacent detections and still be part of the same trip. An elapsed time that is greater than the travel time threshold indicates that one trip has ended and a new trip has begun. The trajectory should be split at this point to represent the two different trips.

The result after clustering and splitting is a dataset consisting of trips, where each trip is a sequence of detections giving the location, date, and time of each detection within each trip. Figure 13 presents an illustration of clustering and splitting a single trajectory. This process is repeated for each of the trajectories in the raw dataset, with the first detection being associated with the origin, and the last detection being associated with the destination.

Raw Data

MAC ID	Location	Date	Time
00:24:9F:D8:F9:E3	15	9/18/2011	1:09:23 PM
00:24:9F:D8:F9:E3	14	9/18/2011	1:12:57 PM
00:24:9F:D8:F9:E3	12	9/18/2011	1:16:12 PM
00:24:9F:D8:F9:E3	12	9/18/2011	1:16:17 PM
00:24:9F:D8:F9:E3	10	9/18/2011	1:19:54 PM
00:24:9F:D8:F9:E3	10	9/18/2011	1:19:59 PM
00:24:9F:D8:F9:E3	8	9/18/2011	1:24:48 PM
00:24:9F:D8:F9:E3	8	9/18/2011	11:47:28 PM
00:24:9F:D8:F9:E3	10	9/18/2011	11:51:02 PM
00:24:9F:D8:F9:E3	12	9/18/2011	11:54:05 PM
00:24:9F:D8:F9:E3	14	9/18/2011	11:56:38 PM
00:24:9F:D8:F9:E3	14	9/18/2011	11:56:43 PM
00:24:9F:D8:F9:E3	15	9/18/2011	11:59:36 PM
00:24:9F:D8:F9:E3	15	9/18/2011	11:59:42 PM
00:24:9F:D8:F9:E3	15	9/18/2011	11:59:47 PM

Clustered Data

MAC ID	Location	Date	Time
00:24:9F:D8:F9:E3	15	9/18/2011	1:09:23 PM
00:24:9F:D8:F9:E3	14	9/18/2011	1:12:57 PM
00:24:9F:D8:F9:E3	12	9/18/2011	1:16:17 PM
00:24:9F:D8:F9:E3	10	9/18/2011	1:19:59 PM
00:24:9F:D8:F9:E3	8	9/18/2011	1:24:48 PM
00:24:9F:D8:F9:E3	8	9/18/2011	11:47:28 PM
00:24:9F:D8:F9:E3	10	9/18/2011	11:51:02 PM
00:24:9F:D8:F9:E3	12	9/18/2011	11:54:05 PM
00:24:9F:D8:F9:E3	14	9/18/2011	11:56:43 PM
00:24:9F:D8:F9:E3	15	9/18/2011	11:59:42 PM

Trip 1

MAC ID	Location	Date	Time
00:24:9F:D8:F9:E3	15	9/18/2011	1:09:23 PM
00:24:9F:D8:F9:E3	14	9/18/2011	1:12:57 PM
00:24:9F:D8:F9:E3	12	9/18/2011	1:16:17 PM
00:24:9F:D8:F9:E3	10	9/18/2011	1:19:59 PM
00:24:9F:D8:F9:E3	8	9/18/2011	1:24:48 PM

Trip 2

MAC ID	Location	Date	Time
00:24:9F:D8:F9:E3	8	9/18/2011	11:47:28 PM
00:24:9F:D8:F9:E3	10	9/18/2011	11:51:02 PM
00:24:9F:D8:F9:E3	12	9/18/2011	11:54:05 PM
00:24:9F:D8:F9:E3	14	9/18/2011	11:56:43 PM
00:24:9F:D8:F9:E3	15	9/18/2011	11:59:42 PM

Figure 13. Illustration of clustering and splitting raw Bluetooth data to devise individual trips.

A final step in the analysis is to factor up the trips into an OD table representing the full population of traffic, as the trips estimated from the raw Bluetooth data represent a fraction of the total traffic, typically 4–15%, as described. The analyst can use a frataring process or iterative proportional fitting to perform this factoring to match the vehicle volume counts at each detector location. The final product of this analysis is a standard OD table, where each detector location is a proxy for an OD zone. The data can also be expressed in select link format.

10.5 Assessment of Data Suitability

10.5.1 Strengths

Bluetooth detector technology is relatively inexpensive to deploy. The process of matching device IDs at various Bluetooth detector stations is relatively straightforward. As a result, this approach to OD estimation is quick and cost-effective, so long as accurate traffic count information for the same highway segment is collected (ideally) simultaneously with the Bluetooth deployment.

The technology is particularly strong when aiming to determine interchange-to-interchange flows along limited access highways. In such cases, the points of access to the highway system are definitive and “noise” from other sources, such as parking lots or general street traffic, can be eliminated or minimized.

The Bluetooth detector data are also passively-detected, which means that they do not burden the traveling public. Furthermore, the MAC IDs associated with Bluetooth devices are linked to a specific device, but not to a specific person or household. As a result, privacy issues are not a concern with this data technology.

10.5.2 Shortcomings

Several Bluetooth data shortcomings need to be considered when using the data for estimating ODs. **Collectively, these issues require substantial data cleaning to produce valid estimates**.

Issues to be considered include the following:

1. Cases where the Bluetooth detectors are situated at locations where vehicles that are not in the study (“noise”) might be detected. This could occur if, for example, a large parking lot were located within detection range. In such a case, the detector will record devices that are in discoverable mode within the parked vehicles, which are not meant to be part of the OD study. Similar unwanted detections can occur near intersections, where cross-street traffic may be recorded but should not be part of the analysis. As noted, data “noise” is minimized in deployments along freeways, which often have broad areas without intersecting traffic or other unwanted detection. Adapting the technology to study areas involving surface streets invite a “noise-rich” environment and complicate the OD estimation process.
2. Cases where the detection range of two Bluetooth detectors overlap, which could lead to the simultaneous recording of the identical MAC ID (from the same device) at two separate locations. This case presents a data problem that can confound the OD analysis. The remedy is relatively straightforward: users must ensure adjacent detectors are not closer than 500 feet, which is well outside the 300-foot (approximate) detection range.

3. Cases where there are multiple Bluetooth devices on one vehicle. This situation could occur in a passenger vehicle with two occupants whose cell phones are in discoverable mode. Another possibility is a transit vehicle with multiple passengers. Data cleaning for this case would need to identify multiple locations where the same pattern of MAC IDs are detected simultaneously.
4. The most concerning issue with Bluetooth technology for OD estimation is the case of missed detections.. A miss could occur because a device is temporarily occluded from the detector while it is within the detection radius. A miss is a case of a specific MAC ID being recorded at stations “A” and “C,” but not at station “B.” The magnitude of this has been estimated to be 5-10% of the total Bluetooth sample, as estimated in cases where duplicate detectors are deployed at the same location. With two detectors at one location, the records of both detectors can be analyzed to determine the fraction of Bluetooth devices that are missed which, in turn, enables the calculation of a correction factor.

11.0 Stated Preference Origin-Destination Data

11.1 Background

Stated preference (SP) data, also often referred to as stated choice or stated response data, are collected to obtain travel choices under hypothetical scenarios. SP data are typically used to obtain choice data for travel alternatives that do not yet exist—such as a new transit mode or a new toll road or managed lane—or to more accurately estimate trade-offs between choice variables that are highly correlated in real situations. An example of the latter is to better estimate models of trade-offs between travel time and travel cost to establish the value of travel time savings (VOTTS, often shortened to VOT). This chapter covers SP methods and their use to model different dimensions of travel behavior. Because SP methods are only occasionally used to model destination choice or origin-destination (OD) patterns, this chapter is less detailed than some of the other chapters.

11.2 Introduction

SP surveys were first used in market research contexts in the 1970s, typically under the label of "conjoint analysis." Their use in travel demand research began in the 1980s and has since grown. SP research has found the following applications in travel demand analysis:

- Studies of route choice preferences for various auto and bicycle trips.
- Studies of preferences across different transit services.
- Choice modeling for the introduction of new tolled highway facilities or new toll pricing systems.
- Mode choice modeling for the introduction of new mode alternatives (e.g., new light rail or high-speed rail).
- Studies of time/cost trade-offs (VOT) for use in project appraisal/benefit-cost analysis.
- Studies of time-of-day pricing and effects on departure time choice.
- Studies of residential choice location decisions.

Studying destination choice has been a relatively rare context for using SP methods, although some examples are given in the following sections.

11.3 Overview of Data Collection

This section provides a brief overview of how SP data are collected. A more complete description can be found in the Transportation Research Board's online *Travel Survey Manual* (Correia & Bradley 1996) and in the textbook by Louviere, et al. (2000).

11.3.1 Survey Questionnaire

A typical SP survey includes the following sections:

- **Establishing the travel choice context**: This involves asking each respondent about a recent relevant trip that he or she made; that trip can then be used as the context for the hypothetical choice options that will be presented. Respondents provide details about their actual trips, such as origin location, destination location, trip purpose, travel party size and composition, time of day, mode used, and actual time or money spent on the trip. Using the context of an actual trip makes the hypothetical choice options more relevant and realistic for the person and helps them to consider any constraints or factors that influenced their actual choice when stating their hypothetical choices. These questions also provide revealed preference (RP) data on actual choices that can often be combined with the SP data in analysis.

- **Presenting the hypothetical travel choice scenarios**: The hypothetical choice options are typically tailored and customized based on the reference trip characteristics reported by the respondents. For example, auto travel times could be offered at levels that differ by prespecified amounts from the reported actual travel time, and auto toll levels and time savings relative to a nontolled option can also be set at prespecified levels depending on trip distance, time of day, and estimated congestion levels. The levels are prespecified and combined into choice alternatives using a statistical experimental design. Orthogonal experimental designs, in which the different choice attributes are varied in a way that removes any statistical correlation between their levels, have been favored in the past. However, more recent emphasis has been on using efficient experimental designs that are optimal for discrete choice model estimation (Rose, et al. 2008).
- **Collecting additional background traveler and household characteristics**: These questions include the typical sociodemographic variables collected in household travel surveys, including income, age, gender, employment status, household size, and vehicle ownership, among others.

Over the years, the format for SP surveys has evolved from primarily paper-and-pencil based surveys to computer-assisted on-site personal interviews (CAPI) and then to online surveys. The transition to computer-based SP experiments facilitated the customization of choice options to each respondent's reference situation. This transition also facilitated more widespread use of graphics and randomization in presenting the choice options. For example, Figure 14 and Figure 15 show examples of paper-based and web-based SP scenarios used to study residential location choices.

The transition from CAPI surveys to online surveys has reduced the cost of conducting SP surveys dramatically while potentially making them accessible to a broader population. The online survey questionnaires can be administered in several ways, and multimethod sampling approaches are typically used to reach different segments of the population. For example, if a survey is targeting users of an existing toll corridor, customers with a transponder can be reached by sending an e-mail invitation, while video tolling customers can be reached by sending a paper-based invitation in the mail. If a facility still offers cash payment options, then cash customers can be recruited by distributing a paper-based invitation at the point of payment. Address-based sampling can also be used to target the general population of a region or residents who live along a specific corridor. Occasionally, on-site personal interviews or telephone-based completion options are still used for the ever-shrinking portion of the population that does not have internet access.

In the next two pages two pairs of neighborhoods are described. After reading the descriptions, please answer the questions that follow. Keep in mind that anything that we do not refer to in the neighborhood choices – such as school quality, public safety, or house size – is exactly the same between the two choices presented.

Neighborhood "W"	***Neighborhood "X"***
To ride public transit, like a bus or train, I have to go **1/4** mile from my house and it comes every **3** minutes.	To ride public transit, like a bus or train, I have to go **10** miles from my house and it comes every **5** minutes.
The distance from my home to the nearest **commercial district**, with things like a retail store, restaurant, movie theater and drug store, is **1** mile. Parking there is **limited, off-street and free**.	The distance from my home to the nearest **commercial district**, with things like a retail store, restaurant, movie theater and drug store, is **2** miles. Parking there is **limited, on-street and costs $3 per hour**.
The housing in my neighborhood consists of **only single family dwellings on 1/8 acre lots**.	The housing in my neighborhood consists of a **mix of townhomes, apartments, condos, and single family dwellings on 1/4 acre lots**.
My one-way commute to work is **1/4** mile.	My one-way commute to work is **20** miles.
Renting or buying a home in this neighborhood is **20% less** compared to your current neighborhood.	Renting or buying a home in this neighborhood is **5% less** compared to your current neighborhood.

33) *Assuming that there are no differences between the neighborhoods apart from the ones we mentioned, which neighborhood do you think you'd rather live in?*

Neighborhood "W" OR ***Neighborhood "X"***

34) *How do you think you'd feel about living in* ***Neighborhood "W"?***

0 1 2 3 4 5 6 7 8 9 10

dislike very much — neutral — like very much

35) *How do you think you'd feel about living in* ***Neighborhood "X"?***

0 1 2 3 4 5 6 7 8 9 10

dislike very much — neutral — like very much

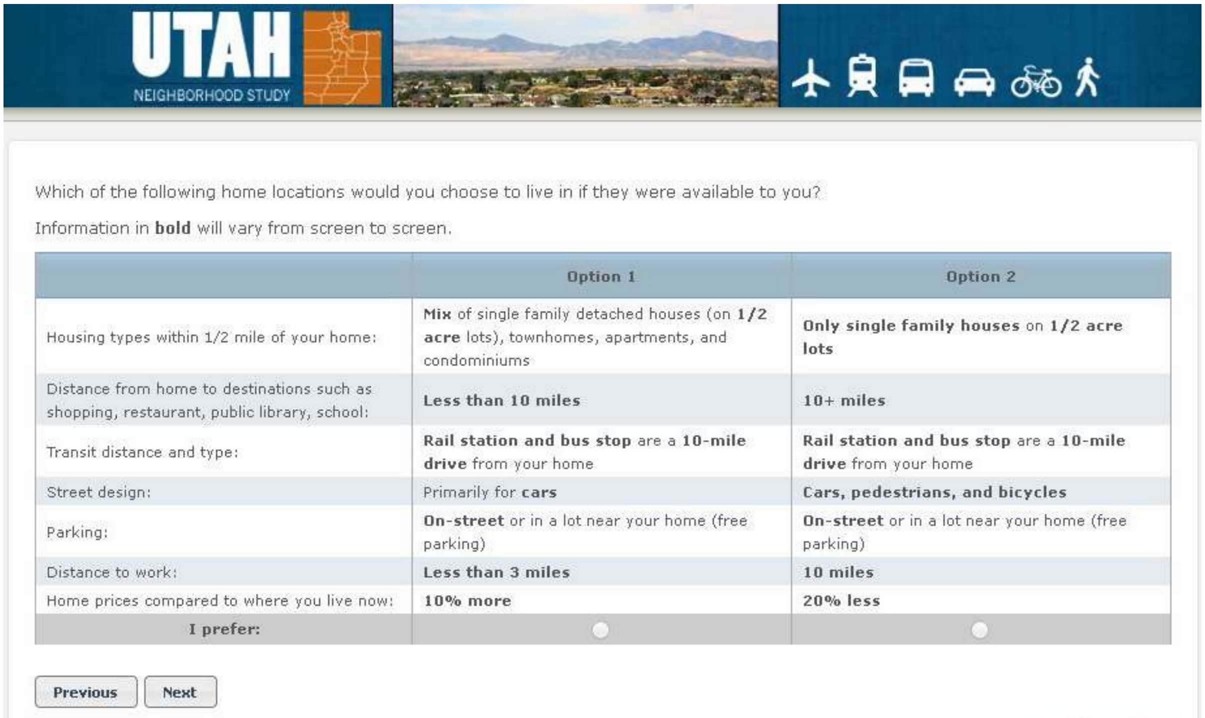
36) *The neighborhood you'd hope to find would be:*

0 1 2 3 4 5 6 7 8 9 10

more like **W** than your current neighborhood — like your current neighborhood — more like **X** than your current neighborhood

10

Figure 14. Paper-based residential location choice SP (from Atlanta SMARTRAQ survey).


UTAH
NEIGHBORHOOD STUDY

Which of the following home locations would you choose to live in if they were available to you?

Information in **bold** will vary from screen to screen.

	Option 1	Option 2
Housing types within 1/2 mile of your home:	**Mix** of single family detached houses (on **1/2 acre** lots), townhomes, apartments, and condominiums	**Only single family houses** on **1/2 acre lots**
Distance from home to destinations such as shopping, restaurant, public library, school:	**Less than 10 miles**	**10+ miles**
Transit distance and type:	**Rail station and bus stop** are a **10-mile drive** from your home	**Rail station and bus stop** are a **10-mile drive** from your home
Street design:	Primarily for **cars**	**Cars, pedestrians, and bicycles**
Parking:	**On-street** or in a lot near your home (free parking)	**On-street** or in a lot near your home (free parking)
Distance to work:	**Less than 3 miles**	**10 miles**
Home prices compared to where you live now:	**10% more**	**20% less**
I prefer:		

Previous Next

Figure 15. Web-based residential location choice SP (from Utah Department of Transportation survey).

11.3.2 Supplemental Data Processing

One of the more attractive features of SP-based modeling is that all the data that the respondents used to make their hypothetical choices are presented during the survey and are already included in the survey dataset. As a result, it is unnecessary to merge the data with external explanatory data from zone-based networks, land-use databases, and so forth. This is a costly and time-consuming step that is usually required for RP modeling.

Some expansion of SP results to the general population may be required, depending on the study context and how the sample was selected. If the SP results are used simply to develop trade-off weights for specific market segments, then expansion and weighting may not be needed, assuming that the market segmentation and model specification includes all of the most important variables that would be used in weighting. In cases where expansion and weighting of SP datasets is carried out, it is typically a less rigorous process than weighting of a full household travel survey, for example.

Some cases exist where OD data may be used in expanding SP data. To illustrate, for a model to represent a specific highway corridor in a toll road study, the SP sample could be expanded to match an estimate of the number of actual car trips using the corridor by origin area/destination area combinations.

11.4 Data Use in Travel Forecasting Models

SP data and SP-based analysis and modeling results can be used in travel forecasting in several ways. Section 11.4.1 discusses a few general approaches. Section 11.4.2 provides more detail on a few application examples related specifically to destination choice and OD patterns.

11.4.1 Direct Application in Sample Enumeration Procedures

One of the most "self-contained" approaches for using SP data and analyses is to apply SP-based travel demand models to the SP survey sample itself and expand and aggregate the results to obtain market-level predictions (e.g., by market segment). This approach has the advantage of not requiring additional survey data on travelers or trips to apply the model. Another advantage is that the software implementation can often be kept simple and easy and fast to run. An example is the Excel-based scenario simulator recently created for an NCRRP SP-based study of mode choice for long-distance trips in the Northeast and Cascades rail corridors (National Academy of Sciences, 2016). In a direct application, external data sources, such as trip OD matrices from aggregate passive data, can be useful in expanding the SP survey sample to be representative of the travel market. In many past applications of this type, there has not been good data of that type available for data expansion, so the growing availability of passive OD data in the future may be valuable.

11.4.2 Transfer of Parameters or Trade-off Ratios to an Existing Model

In some cases, a model application framework may already exist, and SP-based data collection and analysis is used to update the model parameters or the ratios between model parameters (e.g., VOT). A typical example is an existing travel forecasting model to predict use of a new toll facility in a highway corridor. In such a case, the relative values of coefficients for travel time, toll cost, and (if relevant) antitoll bias can be transferred from SP models to the forecasting model.

Although the ratios of coefficients from SP-based models may be appropriate for application in forecasting, the absolute scale of the SP-based coefficients may not be appropriate. In real choice situations, there are typically more unobserved additional factors affecting choices than those that were presented in the SP experiments, so the sensitivity of real-world choices to the SP attributes may be less than was observed in the hypothetical choices. Three main ways exist to address this issue for forecasting:

- Transfer only the ratios of the parameters and use them to combine the effects of multiple variables into a single composite cost, such as combining cost plus time multiplied by VOT into generalized cost. The scale of the composite variable can be estimated by applying the model and calibrating it to real-world choice shares or observed choice elasticities.
- Combine the SP data with RP data on actual choices, with the SP and RP data sharing some or all the same explanatory variables. Bradley and Daly (1997) developed a method for combining SP and RP data in analysis using standard logit estimation software. The RP data can be collected as part of the SP survey (e.g., as the reference trip data) or it can be from a separate survey, such as a household travel diary survey.
- Compare models based on SP and RP datasets to develop a consensus on best application parameters. This is the approach that was used in the SHRP2 C04 project to recommend travel time and cost coefficient functions to use in applied models. That study was based on several different RP and SP datasets of route choice, mode choice, and time-of-day choice from different cities and survey contexts (Parsons Brinckerhoff et al. 2013).

In any of these cases where an existing model is updated with new parameters or ratios of parameters, it is important to calibrate and validate the model against observed choice behavior to the extent possible. The data often used for calibration and validation include household travel diary survey data, traffic facility and screenline count data, and (increasingly) aggregate passive data.

11.4.3 Creating New Model Systems Using SP/RP-Based Models

SP/RP surveys are perhaps the least typical and most resource-intensive type of application. Where these are carried out, joint SP/RP models are estimated from the survey data, and a new model application system is built in which to apply the models. This type of application is needed for projects where no suitable existing model is available, and it is deemed worth the effort to apply the models to a full synthetic population and network zone system rather than simply using the survey data in a sample enumeration. The most common context for such applications is for new proposed choice alternatives for long-distance travel where no suitable long-distance travel model already exists. An example of this application is the model system created to predict demand for the proposed California high-speed rail system. The initial model system was created based on an SP/RP survey (Outwater, et al. 2009), and has since been updated based on further SP/RP surveys and model enhancements. Although the SP choice experiments and models were focused on mode choice, the RP data and full model system predicts destination choice and trip frequency, as influenced by changes in accessibility (mode choice model logsums). Aggregate passive OD data can also be useful in calibrating the destination choice components of such integrated model systems.

11.5 SP Data as Applied to Destination Choice

This section discusses the reasons that SP methods are rarely used to study destination choice behavior. It also discusses the rare contexts where SP has been used to model choice of destination.

11.5.1 Why SP Methods Are Not Well-Suited to Model Destination Choice

In a typical travel demand model system, destination choice is modeled using thousands of different choice alternatives for different geographic zones. Suppose there was interest in destination choice for "shopping." Not only are there hundreds of zones containing possible shopping locations, but each zone can contain dozens of different retail stores of different types, and each of those types of stores could be characterized by many different attributes (e.g., selection and quality of merchandise, prices, availability and friendliness of the staff, adequacy and attractiveness of the retail space, opening hours, parking supply, advertising, special sales, loyalty programs, and many others).

While a retail store operator may be interested in doing an SP study comparing a few similar hypothetical stores along key attributes, such a focused and specific model does not translate to a model system where all types of stores are aggregated into zones and the only data available for the zones are the total number of retail employees and (sometimes) the amount of retail floor space. Even if the model is an AB model that operates on a more detailed spatial scale such as blocks or individual parcels, the amount of data that is available to describe each parcel is nowhere near the amount of detail that one would want to use in an SP study of store destination. Models can reasonably portray the accessibility to different destinations by different modes, but they cannot reasonably portray the attributes of the destinations themselves.

The lack of detail in destination-specific data not only makes SP methods unsuitable, but it also makes destination choice models by far the weakest components of travel demand models. There tends to be much more detailed and relevant data available to describe households, persons, and travel modes and networks than about destination alternatives—particularly in a way that can forecast to provide model inputs for future-year scenarios.

11.5.2 Choice of Residential Locations, Shopping Locations, and Tourist Destinations

Three location choice contexts represent the most common applications of SP methods. The first is choice of shopping destination. As described, this is typically studied from retailers' perspectives ("How do I get more people to shop at my store?"), and not from transportation planners' perspectives ("How do I predict the OD patterns of all the shopping trips in my region?"). The second is residential location choice (as depicted in Figure 14 and Figure 15). Residential choice is somewhat more tractable to study with SP than shopping destination choice, since most people only choose one home in which to live and it is a choice that is considered carefully. Still, there are hundreds of factors that influence residential location choice, and several of them (e.g., schools, amenities, crime, natural surroundings, and housing quality) may be more important than transportation access. As a result, SP-based models of residential choice have rarely (if ever) been used for applied spatial choice forecasting.

A third common context is destination choice for long-distance tourist trips. This is a more tractable and "natural" context for SP choice exercises, since tourist destinations of a specific type (e.g., island/beach resorts, or winter ski resorts) can often be compared on a relatively small set of attributes, some of which can be obtained from land-use data (e.g., supply and prices of hotels, supply and prices of restaurants, types of entertainment), and tourist vacations are sometimes sold as a package deal that includes transportation. Nevertheless, even these studies are more geared toward tourist bureaus and agencies ("How do we attract more visitors?") than they are toward state or national transportation planning ("How do we predict seasonal OD flows for long-distance leisure travel?").

Because none of these three typical location choice SP contexts are used for spatial planning or forecasting, the potential for using aggregate passive OD data is limited, at least in the short term. Over time, if the passive data were accurate enough to locate trip ends to stores, and parcel land-use databases were accurate regarding attributes of the retail establishments on each parcel, then it may be feasible to apply more detailed models of shopping location choice and to use SP surveys to help estimate the models.

11.5.3 Studies of Area Pricing or Cordon Pricing

One type of transportation policy that affects destination choice in a way that is amenable to presentation in SP experiments is area-based or cordon-based pricing policies. SP surveys were used to study potential time-of-day-based cordon-pricing options in downtown San Francisco (San Francisco County Transportation Authority, 2010). The SP experiments offered assorted options to avoid paying the cordon charge (or pay a lower cordon charge):

- Switch to transit instead of driving.
- Drive at a different time of day.
- Choose a destination outside of the cordoned area.
- Skip the trip.

This SP context offered destinations that are defined only by whether they are inside or outside the cordon, without defining exactly what destination outside the cordon would be selected instead. In this case, the destination choice context is not specific enough to calibrate or validate using existing choice data where there is no pricing cordon. Also in this case, the cordon pricing was never implemented, so no before-and-after study could be conducted. If it were implemented, then comparing aggregate passive OD data for different weeks before and after the policy enactment would effectively validate/calibrate the SP-based models, in terms of both the destination-switching and time-of-day switching responses to the policy. This emphasizes the point that SP studies, because of their hypothetical nature, are pseudodynamic. In other words, when they include nonexistent alternatives, they require longitudinal before-and-after data for true real-world validation. Passive data sources offer a relatively low-cost method of obtaining consistent aggregate data for different points in time.

11.5.4 Studies of the Relationship Between Destination Choice and Other Travel Choices

A recent study that included destination choice studied the structural relationship between destination choice and toll versus nontoll route choice for auto trips in Nigeria (Davidson et al. 2014). To simplify the SP experiment, respondents only had to consider two possible destinations—one where they want to travel, and another that they are familiar with and where they could carry out the same activity, but that may require a different travel time and cost to get there. As the authors discuss, this simplification does introduce possible self-selection effects and major differences from how destination choice models are defined in practice. The authors' objective, however, was to investigate the hierarchy between destination choice and route choice in a nested model structure. The results indicate that toll/nontoll route choice should be nested below destination choice in contexts where the destination activity is transferable from one location to another. This result supports the nesting structure that is typically used in forecasting models (route choice below destination choice). Using longitudinal passive aggregate data on OD patterns and route choice patterns before and after a new toll facility (or a major toll pricing change) is introduced could provide useful real-world data for validating similar SP-based findings.

12.0 Traffic Counts in the Origin-Destination Estimation Process

12.1 Background

Many technical approaches to estimating origin-destination travel rely at least partially on traffic counts for sample expansion and for validation. For example, origin-destination (OD) studies utilizing Bluetooth detectors sample a small fraction (e.g., 5-10%) of the total vehicle population. For this approach, having accurate traffic counts conducted simultaneously with the Bluetooth deployment provides the analyst with the expansion factors needed to estimate an OD matrix. OD estimation processes using passively collected sources such as GPS traces from navigational devices or smartphone apps similarly require traffic count information to estimate the percentage of the travel demand their sample represents. In addition, traffic counts are essential for overcoming systematic biases in passively collected data related to coverage, trip length, or trip duration. Additional information on using traffic counts for sample expansion and validation can be found in the FHWA "How-To series, "How-To: Develop Big Data Driven Demand for Traffic Forecasting." This source also features a detailed description of one of the most widely used OD estimation methods, OD Matrix Estimation (ODME) from counts and other methods of using counts together with OD data to improve estimates of OD patterns.

12.2 Introduction and Overview of Data Collection

As discussed, traffic counts are widely used in most OD estimation processes to expand estimates and to provide validation datasets. FHWA's *Traffic Monitoring Guide* provides a detailed background into traffic count technologies, programs, and sources and magnitude of error.[12] A high-level overview is provided here.

Traffic counts can be collected in many ways, ranging from the manual pencil and paper approach to counts obtained with pneumatic tubes, count boards, radar, video recognition and other methods. A final count volume is usually associated with a highway segment, or with an intersection movement, by time of day. Most state departments of transportation administer traffic count programs that are designed to capture hourly, daily, and seasonal variation across many functional class highways. These count programs typically combine continuous count stations with short-duration (e.g., 2-day, 2-week) counts.

Errors can occur during traffic counts for many reasons including human error and machine-recognition errors. An example of the latter is when weather occludes the actuation area of a video counter and vehicles are miscounted. A common error with pneumatic tubes is when the tubes break or break free from the roadway surface. Counts conducted through inductance technology are vulnerable to loop breakage. A set of traffic counts that is assembled for a travel model may have other inconsistences, such as the following:

- Collection at different times/on different days.
- Collection at suboptimal locations.

[12] Federal Highway Administration Traffic Monitoring Guide.

- Effect from special events like incidents, construction, or weather.

Despite these limitations, traffic counts, whether conducted manually or by machine, or by a combination of both, are a foundation of travel model calibration, and of the OD validation process for many traditional and emerging approaches. Traffic data collection approaches are described above, but more detail is provided in FHWA's *Traffic Monitoring Guide*.[13] FHWA's *Traffic Detector Handbook* provides detail on the technologies used for counting traffic, their strengths, and limitations.

12.3 Preparing Origin-Destination Estimates Using Traffic Counts

For simple cases, particularly ones involving highways with well-defined access points, a set of traffic counts might be sufficient to confidently estimate the origin-destination pattern. An example of this is shown in Figure 16.

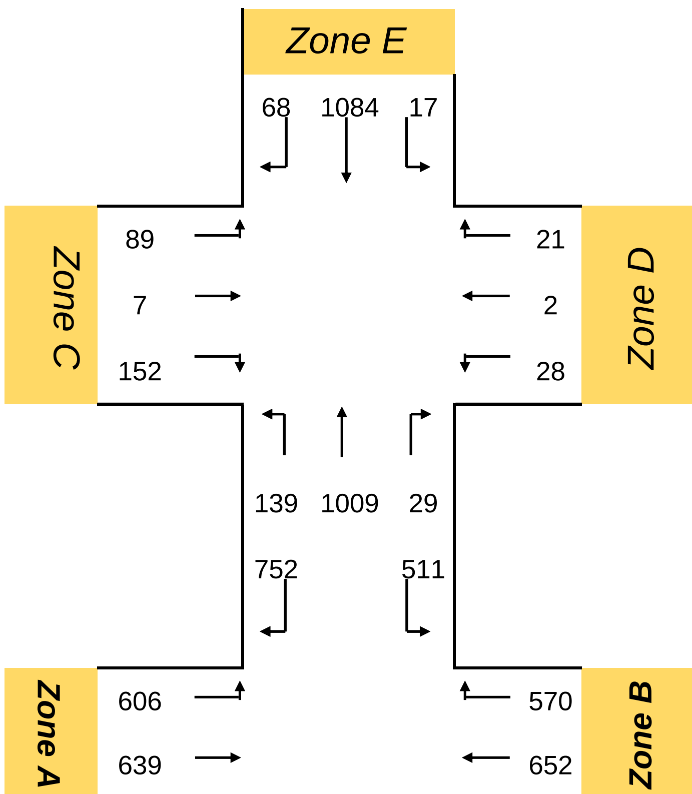


Figure 16. Illustration of simple study area for estimating OD matrix.

[13] Federal Highway Administration Traffic Monitoring Guide

For this simple case, the OD matrix can be estimated by assuming simple proportionality of the turning movements, with the resulting matrix shown in Table 2. (Although it is worth noting that this assumption may or may not obtain.)

Table 2. OD matrix resulting from the traffic counts in Figure 16.

	A	B	C	D	E
A	--	639	71	15	519
B	652	--	67	14	489
C	90	61	--	7	89
D	17	11	2	--	21
E	645	438	68	17	--

In practice, travel models contain significantly greater complexity, with many more traffic analysis zones (TAZs). Simple proportionality is not sufficient for accurately estimating an OD matrix in these cases.

12.3.1 OD Matrix Estimation

One of the most common methods for estimating an OD matrix from traffic counts is ODME. Several ODME algorithms are in use and these algorithms have diverse properties that can produce significantly different results. ODME estimation methods have been covered in detail in other reports and studies, such as NCHRP Report 765: *Analytical Travel Forecasting Approaches for Project-Level Planning and Design*. This chapter will not cover the more sophisticated ODME processes.

The overall ODME estimation process involves minimizing errors compared to traffic counts. Traffic counts reveal information about OD patterns, but are not sufficient, by themselves, for developing an accurate estimate. A key challenge of this approach is that, mathematically, there are many more OD flows than traffic counts. Consequently, the problem is statistically underdetermined, with more than one possible set of OD results that successfully satisfy the constraints imposed by the traffic counts. For this reason, more sophisticated ODME procedures involve additional constraints beyond error minimization versus counts. Another important feature of ODME is its dependence on a well-validated network assignment model. Any errors in the network assignment model used in the ODME process could cause corresponding errors in the estimated ODs.

The ODME process starts by adjusting and iterating from an initial seed matrix. One can start a seed matrix using traffic counts, with proportionally allocated origins and destinations, like the process described above. However, producing a final OD matrix purely by minimizing errors versus traffic counts can introduce significant estimation errors.

Figure 17 shows an example of a limited access highway system, with traffic count volumes associated with highway segments and ramps. In this case, each on- and off-ramp can be viewed as an origin or destination zone. The analyst can calculate total entering and exiting volumes for each zone as a start to developing the seed matrix. Other seed matrix starting points could be generated by the following:

- Subarea analysis from a calibrated regional model.
- Rough gravity model with overall zone OD marginal sums dictated by either counts (preferably) or trip generation calculations.
- OD pairs that have known values from turning movement count data (zone pairs at the edge of a model extent) can be hard coded from the count into the seed matrix.
- OD information from another method, such as a Bluetooth detector deployment.

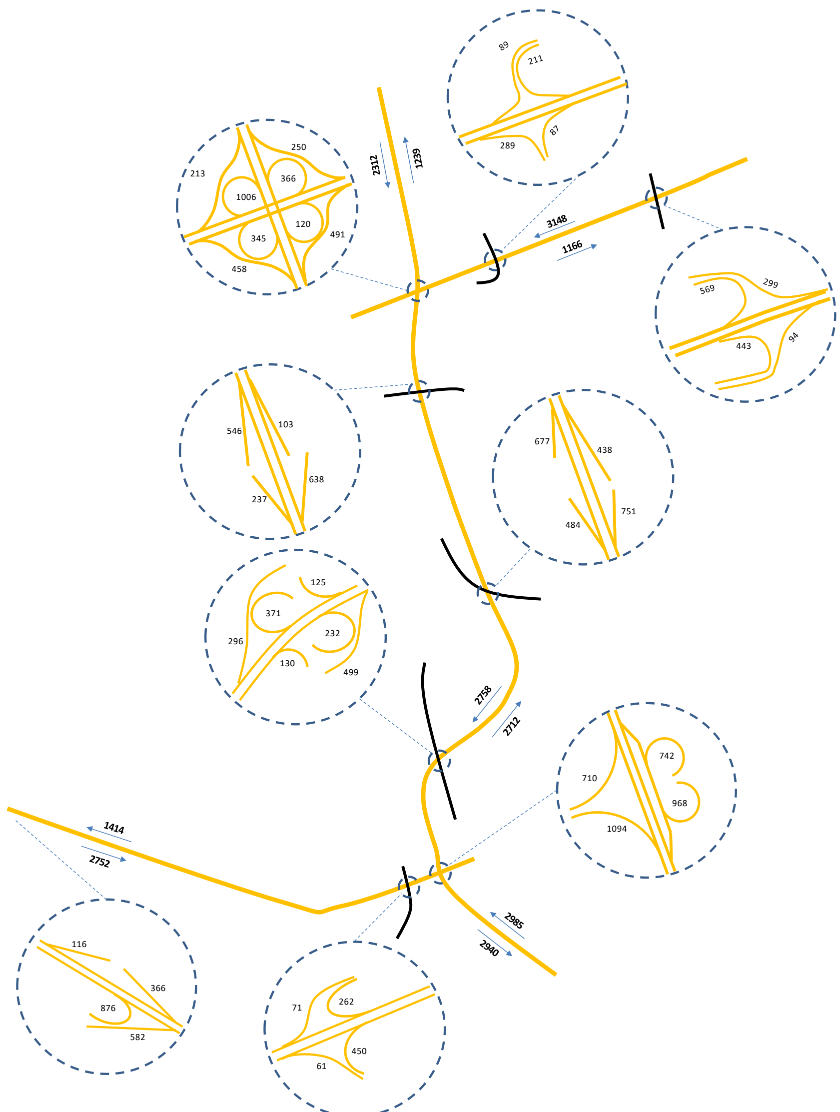


Figure 17. Illustration of more complex study area for OD estimation.

Introducing additional constraints to the ODME algorithm using information from other datasets or models can improve the accuracy of the final OD matrix. Additional constraints might include information about OD flows, or limitations on how much the final matrix varies from the seed matrix. Other ODME algorithms allow an analyst to define the maximum positive and negative changes for each zone pair. In general, the most limited adjustments capable of producing good agreement with traffic counts are desirable. A more complete description of the ODME alternatives and their strengths and weaknesses is described in Volume 4 of this series, "Estimating Origin-Destination Data Using Data Fusion: A Proof of Concept."

13.0 Summary and Conclusion

New data sources can help address existing gaps in traditional data sources. Volumes 2, 3, and 4 of this series investigate how Big Data sources can be used to complement or augment. Generally, a key insight from this work is that a single source of available data—either survey or passive data—is not adequate for developing accurate origin-destination (OD) estimates. However, when a traditional method is paired with an OD dataset developed from Big Data sources, together with appropriate control data for expansion, the strengths of both the traditional and Big Data approaches can be leveraged for greater accuracy. Using only one approach or one dataset, from traditional or Big Data sources, should be considered inadequate for OD estimation.

14.0 References

Bellafante, G. (2017). “How Much Tourism Is Too Much?” *The New York Times*.

Bradley, M. and Daly, A. (1997). “Estimation of Logit Choice Models Using Mixed Stated Preference and Revealed Preference Information.” In *Understanding Travel Behavior in an Era of Urban Change*. Emerald Group Publishing Limited, United Kingdom, pp. 209–232.

Cambridge Systematics. (2013). *Guidebook for Developing Subnational Commodity Flow Data*, NCHRP Report 26, Transportation Research Board, Washington, DC.

Cambridge Systematics. (2017). *Effect of MTC Express Lanes on Interregional Travel*, Report, Metropolitan Transportation Commission, Oakland, CA.

Chu, X. and Polzin, S.E. (2017). *An Application for Measuring Vehicle Travel by Visitors*, University of South Florida, Tampa, FL.

Correia, G. and Bradley, M. (1996). “Stated-Preference Surveys.” In *The On-Line Travel Survey Manual: A Dynamic Document for Transportation Professionals, Chapter 13*. Provided by the Members and Friends of the Transportation Research Board’s Travel Survey Methods Committee.

Davidson, P, Teye, C., and Culley, R. (2014). “Developing a Successful Stated Preference Methodology for Determining Destination Choice Coefficients and Using it to Investigate its Empirical Structural Relationship with Toll Route Choice.” *Transportation Research Procedia, 1*, pp. 36–48, Elsevier B.V., Amsterdam, The Netherlands.

Farnsworth, S.P., and Hard, E. 2013. *2013 Metropolitan Area Planning Agency External Travel Survey*, Summary Report.

Green, A.S., Kutzbach, M.J., and Vilhuber, L. (2017). *Two Perspectives on Commuting: A Comparison of Home to Work Flows Across Job-Linked Survey and Administrative Files*, Report No. CES 17-34, U.S. Census Bureau, Center for Economic Studies, Washington, DC.

Hard, E., Chigoy, B., Songchitruksa, P., Farnsworth, S., Borchardt, D., and Green, L. 2016. *Synopsis of New Methods and Technologies to Collect Origin-Destination (OD) Data*, Report No. FHWA-HEP—16-083, Federal Highway Administration, Washington, DC.

Jeon, K., Maneva, P., and Livshits, V. (2014). *“Developing and Validating Regional Travel Forecasting Models with CTPP Data: MAG Experience*.” Presented at the CTPP Webinar and Discussion.

Louviere, J.J., Hensher, D.A., and Swait, J.D. (2000). *Stated Choice Methods: Analysis and Applications*. Cambridge University Press, Cambridge, United Kingdom.

Outwater, M., Tierney, K., Bradley, M., Sall, E., Kuppam, A., and Modugula, V. (2010). “California Statewide Model for High-Speed Rail.” *Journal of Choice Modeling*, *3*, pp. 58–83, Elsevier Ltd., Amsterdam, The Netherlands.

Parsons Brinckerhoff, Northwestern University, Mark Bradley Research & Consulting, University of California at Irvine, Resource Systems Group, University of Texas at Austin, Frank Koppelman, and GeoStats. (2013). *Improving Our Understanding of How Highway Congestion* *and Pricing Affect Travel Demand*, SHRP2 Capacity Research, Report S2-C04-RW-1, Transportation Research Board. Washington, DC.

Rose, J. M., Bliemer, M.C.J., Hensher, D.A., and Collins, A.T. (2008). "Designing efficient stated choice experiments in the presence of reference alternatives." *Transportation Research Part B: Methodological*, *42*, pp. 395–406, Elsevier Ltd., Amsterdam, The Netherlands.

RSG. (2016). *Intercity Passenger Rail in the Context of Dynamic Travel Markets*, NCRRP Report 4, Transportation Research Board, Washington, DC.

San Francisco County Transportation Authority. (2010). *San Francisco Mobility, Access, and Pricing Study*, Final Report, San Francisco, CA.

U.S. Census Bureau. (2008). *A Compass for Understanding and Using American Community Survey Data: What General Data Users Need to Know*, U.S. Department of Commerce, Economics and Statistics Administration.

NOTICE

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-19-043

U.S. Department of Transportation
Federal Highway Administration