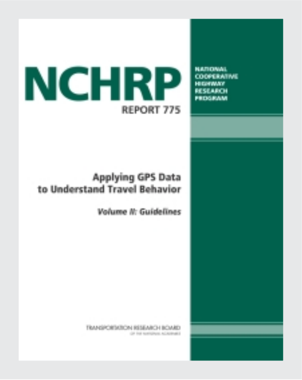


NCHRP
REPORT 775

NATIONAL COOPERATIVE HIGHWAY RESEARCH PROGRAM

Applying GPS Data to Understand Travel Behavior

Volume II: Guidelines

TRANSPORTATION RESEARCH BOARD

Applying GPS Data to Understand Travel Behavior, Volume II: Guidelines

DETAILS

216 pages | 8.5 x 11 | PAPERBACK
ISBN 978-0-309-37519-1 | DOI 10.17226/23436

AUTHORS

Wolf, Jean; Bachman, William; Oliveira, Marcelo Simas; Auld, Joshua; Mohammadian, Abolfazl (Kouros); and Vovsha, Peter

NATIONAL COOPERATIVE HIGHWAY RESEARCH PROGRAM

NCHRP REPORT 775

Applying GPS Data to Understand Travel Behavior

Volume II: Guidelines

Jean Wolf
WESTAT | GEOSTATS SERVICES
Atlanta, GA

William Bachman
WESTAT | GEOSTATS SERVICES
Atlanta, GA

Marcelo Oliveira
WESTAT | GEOSTATS SERVICES
Atlanta, GA

Joshua Auld
UNIVERSITY OF ILLINOIS, CHICAGO
Chicago, IL

Abolfazl (Kouros) Mohammadian
UNIVERSITY OF ILLINOIS, CHICAGO
Chicago, IL

Peter Vovsha
PARSONS BRINCKERHOFF, INC.
New York, NY

Johanna Zmud
RAND CORPORATION
Washington, D.C.

Subscriber Categories
Highways • Data and Information Technology • Planning and Forecasting

Research sponsored by the American Association of State Highway and Transportation Officials
in cooperation with the Federal Highway Administration

TRANSPORTATION RESEARCH BOARD

WASHINGTON, D.C.
2014
www.TRB.org

NATIONAL COOPERATIVE HIGHWAY RESEARCH PROGRAM

Systematic, well-designed research provides the most effective approach to the solution of many problems facing highway administrators and engineers. Often, highway problems are of local interest and can best be studied by highway departments individually or in cooperation with their state universities and others. However, the accelerating growth of highway transportation develops increasingly complex problems of wide interest to highway authorities. These problems are best studied through a coordinated program of cooperative research.

In recognition of these needs, the highway administrators of the American Association of State Highway and Transportation Officials initiated in 1962 an objective national highway research program employing modern scientific techniques. This program is supported on a continuing basis by funds from participating member states of the Association and it receives the full cooperation and support of the Federal Highway Administration, United States Department of Transportation.

The Transportation Research Board of the National Academies was requested by the Association to administer the research program because of the Board's recognized objectivity and understanding of modern research practices. The Board is uniquely suited for this purpose as it maintains an extensive committee structure from which authorities on any highway transportation subject may be drawn; it possesses avenues of communications and cooperation with federal, state and local governmental agencies, universities, and industry; its relationship to the National Research Council is an insurance of objectivity; it maintains a full-time research correlation staff of specialists in highway transportation matters to bring the findings of research directly to those who are in a position to use them.

The program is developed on the basis of research needs identified by chief administrators of the highway and transportation departments and by committees of AASHTO. Each year, specific areas of research needs to be included in the program are proposed to the National Research Council and the Board by the American Association of State Highway and Transportation Officials. Research projects to fulfill these needs are defined by the Board, and qualified research agencies are selected from those that have submitted proposals. Administration and surveillance of research contracts are the responsibilities of the National Research Council and the Transportation Research Board.

The needs for highway research are many, and the National Cooperative Highway Research Program can make significant contributions to the solution of highway transportation problems of mutual concern to many responsible groups. The program, however, is intended to complement rather than to substitute for or duplicate other highway research programs.

NCHRP REPORT 775

Project 8-89
ISSN 0077-5614
ISBN 978-0-309-28402-8
Library of Congress Control Number 2014939657

COPYRIGHT INFORMATION

NOTICE

The project that is the subject of this report was a part of the National Cooperative Highway Research Program, conducted by the Transportation Research Board with the approval of the Governing Board of the National Research Council.

The members of the technical panel selected to monitor this project and to review this report were chosen for their special competencies and with regard for appropriate balance. The report was reviewed by the technical panel and accepted for publication according to procedures established and overseen by the Transportation Research Board and approved by the Governing Board of the National Research Council.

The opinions and conclusions expressed or implied in this report are those of the researchers who performed the research and are not necessarily those of the Transportation Research Board, the National Research Council, or the program sponsors.

The Transportation Research Board of the National Academies, the National Research Council, and the sponsors of the National Cooperative Highway Research Program do not endorse products or manufacturers. Trade or manufacturers' names appear herein solely because they are considered essential to the object of the report.

Published reports of the

NATIONAL COOPERATIVE HIGHWAY RESEARCH PROGRAM

are available from:

Transportation Research Board
Business Office
500 Fifth Street, NW
Washington, DC 20001

and can be ordered through the Internet at:

http://www.national-academies.org/trb/bookstore

Printed in the United States of America

COOPERATIVE RESEARCH PROGRAMS

CRP STAFF FOR NCHRP REPORT 775

Christopher W. Jenks, *Director, Cooperative Research Programs*
Christopher Hedges, *Manager, National Cooperative Highway Research Program*
Nanda Srinivasan, *Senior Program Officer*
Charlotte Thomas, *Senior Program Assistant*
Eileen P. Delaney, *Director of Publications*
Doug English, *Editor*

NCHRP PROJECT 8-89 PANEL
Field of Transportation Planning—Area of Forecasting

Rebekah Anderson, *Ohio DOT, Columbus, OH* (Chair)
Raj Bridgelall, *Upper Great Plains Transportation Institute, Fargo, ND*
Ju-Yin Chen, *Virginia DOT, Richmond, VA*
Christopher M. Puchalsky, *Delaware Valley Regional Planning Commission, Philadelphia, PA*
Richard Roisman, *Metropolitan Washington Council of Governments, Washington, DC*
Elizabeth Sall, *San Francisco County Transportation Authority, San Francisco, CA*
Morteza Tadayon, *Maryland DOT, Baltimore, MD*
Kermit Wies, *Chicago Metropolitan Agency for Planning, Chicago, IL*
Shuming Yan, *Washington State DOT, Seattle, WA*
Elaine R. Murakami, *FHWA Liaison*
Kimberly Fisher, *TRB Liaison*

CONTENTS

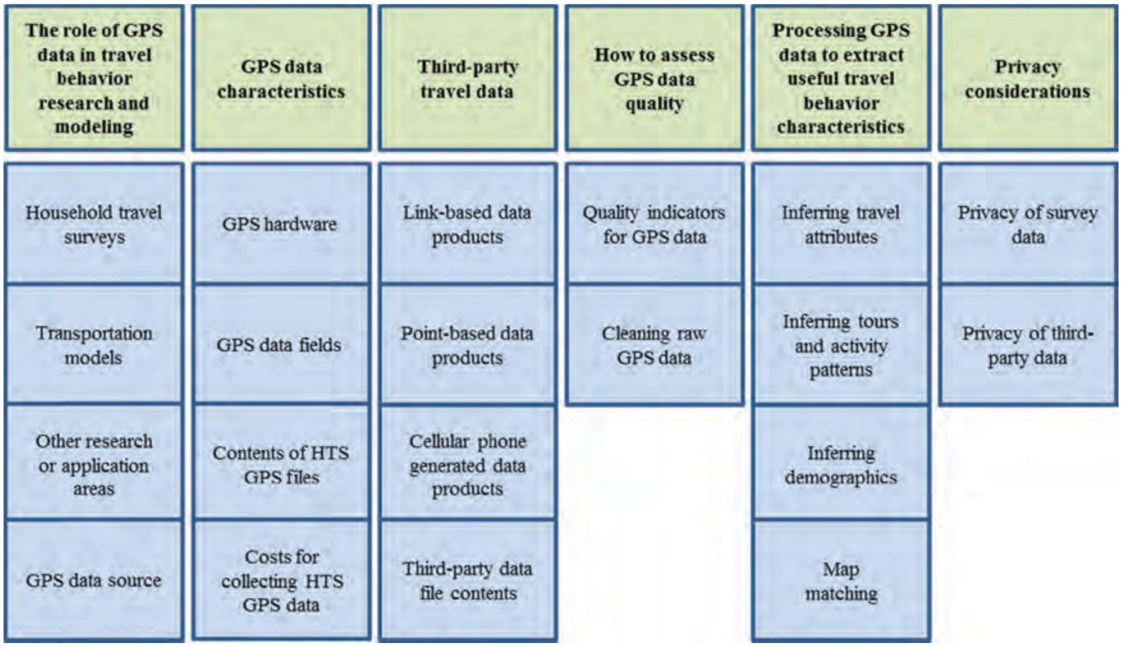
Note: Many of the photographs, figures, and tables in this report have been converted from color to grayscale for printing. The electronic version of the report (posted on the Web at www.trb.org) retains the color versions.

Introduction

The guidelines provided in this document describe the capabilities of Global Positioning System (GPS) data as they relate to transportation planning and specifically to improving the understanding of personal travel behavior. This guidance is intended for transportation planners and researchers who are considering the use of GPS data for travel behavior analysis and would like to jump-start their efforts with the foundational knowledge and recommendations described in this research. As part of this guidance, examples of successful GPS data processing techniques and information on key processing and analysis issues are provided. These steps can be followed regardless of the GPS data source (i.e., travel survey, other field data collection, or data purchased from a third party).

It should be noted that this document is not intended to be used as a step-by-step manual. Instead, it has been developed to facilitate the use of GPS data for transportation planning purposes in an ever-changing environment where new technologies, techniques, and data products are constantly being introduced. Figure 1 depicts the organization of topics that are included in this guidance document.

The creation of this guidance document follows the completion of two core tasks: (1) a comprehensive literature review and industry assessment, and (2) testing and implementation of the most promising GPS data processing techniques; details of these tasks can be found in Volume I of this report. These first two tasks provided in-depth analyses and assessments that are briefly referenced here; readers interested in more background information and details of the testing methods and findings should refer to Volume I. Furthermore, GPS technology, data availability, and recommended uses are highly dynamic topics. To counter this and ensure that this guidance maintains relevance, the discussion contained herein targets key concepts, techniques, and findings that are not tied to specific GPS data or specific products.



The role of GPS data in travel behavior research and modeling	GPS data characteristics	Third-party travel data	How to assess GPS data quality	Processing GPS data to extract useful travel behavior characteristics	Privacy considerations
Household travel surveys	GPS hardware	Link-based data products	Quality indicators for GPS data	Inferring travel attributes	Privacy of survey data
Transportation models	GPS data fields	Point-based data products	Cleaning raw GPS data	Inferring tours and activity patterns	Privacy of third-party data
Other research or application areas	Contents of HTS GPS files	Cellular phone generated data products		Inferring demographics	
GPS data source	Costs for collecting HTS GPS data	Third-party data file contents		Map matching	

Figure 1. Guidance document organization and topic areas.

The Roles of GPS Data in Travel Behavior Research and Modeling

The Global Positioning System was designed by the U.S. Department of Defense in the 1970s as a navigation aid for the U.S. military; it became fully operational in 1995. Nonmilitary and consumer uses of GPS were envisioned from the start. Beginning with a 1996 trial of GPS data collection for transportation planning uses sponsored by the FHWA in Lexington, Kentucky (Battelle Memorial Institute 1997), the use of GPS technology has become increasingly widespread within the transportation field. This can be credited to ever-increasing computer power and storage capacity and the ensuing proofs of concept using GPS data to advance a range of research topics, including travel behavior, route choice, destination choice, and travel speed. A GPS approach for transportation planning data needs offers the benefits of highly precise location, time, and speed data obtained via passive and objective methods.

The promise of these benefits has allowed researchers and practitioners in transportation planning to explore and implement a wide array of practical applications. Of particular interest to travel behavior researchers are GPS applications in traditional household travel surveys (HTSs) and the potential uses of consumer-based travel behavior data products. Other transportation planning areas, such as congestion management, freight/goods movement, real-time traffic information, and asset management, have also used GPS in some manner but are not discussed in detail in this report.

The use of GPS data in transportation research is an increasingly important tool in a researcher's toolkit. Survey tasks that were once accomplished with pencil and paper are now done passively and accurately through the use of GPS technology and systematic data management practices. As this knowledge is passed on to practitioners and they experiment with this technology for the first time, it is important to be familiar with its accepted uses, limitations, and pitfalls. An overview of some of the more promising transportation planning application areas is presented next to provide context for the following guidance provided on GPS data sources and processing methods. Readers interested in more details of these application areas should refer to Volume I. In addition, researchers curious about how GPS works should refer to "Defining GPS and GPS Capabilities" (Wolf 2004).

The next sections address the following topics:

Household Travel Surveys	Reviews the roles of GPS in modern household travel surveys
Transportation Models	Describes the various roles of GPS data in transportation model development and validation

Other Research or Application Areas	Lists additional examples of GPS in transportation planning and research
GPS Data Source	Describes sources of GPS data available to transportation researchers

Household Travel Surveys

In household travel surveys, which are typically conducted as two-stage surveys with sociodemographic data collected first and detailed travel behavior reported next, GPS data collection helps minimize the burden participants experience when reporting their travel. It is a commonly observed outcome of surveys using handwritten travel diaries that response rates are inversely related to respondent burden (National Research Center Inc. 2007). Automating much of the process of reporting travel through the use of passive GPS data collection (initially made possible by GPS data loggers) makes completion of the travel survey more likely, more accurate and complete, and, potentially, cheaper through the use of alternate GPS technologies already in use (including smartphones) by targeted populations and the benefits of multiday data to possibly reduce required sample sizes.

In addition to the information directly recorded in the GPS data set (e.g., start and end time of each trip and location coordinates for each trip start, end, and route), it is possible in some cases to infer other trip details such as travel mode and trip purpose. If GPS data from multiple members of the household are available, it becomes possible to infer partial and full joint trips. It is also possible to better understand resource allocations in the household (e.g., who will take the car). Guidelines for processing GPS data to obtain these travel behavior details are provided in this document.

The list that follows provides a summary of reported advantages that GPS data collection offers within the context of household travel surveys:

- Minimizes respondent burden via passive data collection and imputation methods.
- Increases the likelihood that respondent will fully complete the survey.
- Enhances the spatial and temporal accuracy of the travel behavior data collected.
- Supports derivation or imputation of travel attributes, including travel mode, route, activity location, trip and activity durations, and trip purpose for all trips captured.
- Allows for identification of partial and full joint trips.
- Has the potential to reduce required sample size by capturing multiday travel data.

There are several good examples of household travel surveys conducted around the world that have implemented GPS data collection methods. Volume I of this report discusses these studies, and their trends over time, in more detail. Transportation planners and researchers are encouraged to review Volume I as well as project reports and publications from the following applied studies:

- The 1996 FHWA Lexington Pilot Study: this is the landmark study for considering GPS use in travel surveys.
- The Denver, Atlanta, and California statewide travel surveys (2008–2012): these studies are examples of large-scale surveys that have a GPS subsample paired with a standard diary survey.

- The New York City (New York, New Jersey, and Connecticut) Regional Travel Survey (2010–2012): This survey of more than 19,000 households included a 10% GPS-based prompted-recall sample.
- The Jerusalem, Israel, Travel Habits Survey (conducted in 2010) used a GPS-based prompted-recall approach conducted via face-to-face interviews for 100% of more than 8,000 households that participated.
- The Ohio approach to GPS-based household travel surveys: In 2009 and 2012, Ohio implemented two large-scale, GPS-only household travel surveys. These surveys, conducted in Cincinnati and Cleveland, included a large GPS-only subsample that used GPS data loggers as the only means for reporting travel.

There is also a growing interest in gathering GPS-based travel survey information by utilizing the GPS equipment that many people already carry in the form of smartphones or navigation devices. Planners and researchers interested in the approach should review the CycleTracks app and its success in gathering bicycle data. The project and application were developed by the San Francisco Transportation Authority and have been used as a model for several similar or developing applications. Volume I of the report contains more details on this evolving research area.

Modeling Support

The decision-making demands on applied transportation models are requiring an ever-increasing level of complexity to estimate transportation policy impacts beyond capacity expansion (Cambridge Systematics et al. 2012). The increasing complexity of models and their planning roles require higher-quality data to identify travel behavior and transportation system existing conditions. GPS technology has been targeted as an important tool for collecting the quality data needed in today's models. Over the last decade, the following uses of GPS data have been applied in transportation planning model development:

- Generating trip rate correction factors.
- Identifying activity schedules.
- Exploring activity interactions within a household and within larger social networks.
- Identifying activity locations.
- Identifying route choice and mode choice preferences.
- Exploring variability and pattern formation in activity-travel patterns.
- Identifying baseline network roads and conditions.
- Evaluating bike/pedestrian travel behavior.
- Validating travel demand models.
- Identifying trip purpose and activity type.

These listed roles for GPS data in modeling support have varying degrees of successful application, and readers should review Volume I for more details. In addition to GPS data collected as part of a travel survey or probe vehicle study, many transportation modeling efforts are using or evaluating data collected from various consumer devices and sold as third-party data products. In 2011, 15 state departments of transportation (DOTs) were using private data to supplement their needs, including data from the providers INRIX, SpeedInfo, and NAVTEQ (TTI and Lee Engineering 2011). Furthermore, GPS probe vehicles offer real-time information on flows and speeds, but without the user information needed for many forms of behavior modeling. Finally, while not a GPS-based solution, cell phone triangulation methods for location determination have also been used to capture population movement patterns; private firms market this type of data to state DOTs and regional planning agencies as both a primary source of travel behavior data and as supplemental data that can be used to validate travel demand model outputs.

The intrinsic capabilities of GPS data (i.e., detailed spatial and temporal traces of individuals or vehicles) allow for research into travel behavior beyond the applied uses mentioned previously. The following examples highlight potential research-level efforts that could be pursued to enhance or improve the understanding of travel behavior as applied in transportation models:

- Analysis of visitor/tourist travel behavior.
- Evaluation of commercial vehicle and/or taxi vehicle activity patterns.
- Statistical analysis of GPS traces for overall trends and motivations.
- Identification of overall travel preferences of participants.
- Comparison of routes against a labeled choice set for visited origin–destination pairs.

There are a number of resources available to planners and researchers that discuss the application of GPS in transportation models; these are provided in Volume I. Many potential applications have yet to be documented in an applied setting but can be found in several research efforts. The following sections highlight two select opportunities.

Activity and Trip End Locations

For any travel model, whether a four-step or activity-based model (ABM), a trip record with unknown or incorrect destination location(s) is unusable for most sub-models. Rounding and other errors in trip departure and arrival times are less critical for four-step models since they operate with broad 3- to 4-hour time periods. However, advanced ABMs are extremely sensitive to both spatial and temporal inconsistencies. They operate with tours rather than trips, and having a data item missing on one of the trips frequently results in discarding the entire tour. Differences between travel-diary–based and GPS-based travel time data and how they compare with modeled estimates based on the same origin–destination pairs were analyzed by GeoStats for the three regions in California that participated in the GPS component of the 2001 statewide travel survey (Wolf, Oliveira, and Thompson 2003). This analysis revealed that diary-based reporting tends to significantly overestimate travel times when compared with GPS-derived trips and modeled travel times.

GPS technology can be used to ensure a consistent daily chain of trips and activities for a person because both the spatial and temporal aspects are present in the GPS stream with a high level of detail for routes and modes. In particular, for auto trips, such data as toll facilities or managed lanes used on the trip can be automatically retrieved. For transit trips, the GPS stream provides information about the sequence of all access and transit line segments (including exact boarding, alighting, and transfer points). The GPS stream also clearly identifies the parking location for both auto and transit trips with auto access or egress. Rail transit trips can be captured at the points where travelers enter and exit aboveground stations. Underground connections are not captured explicitly but can sometimes be inferred by analyzing transit schedule data.

Guidance regarding methods for capturing these trip details can be found in the section of this report titled “Processing GPS Data.”

Baseline Model Speeds

Baseline model speeds have historically been developed from tables that identify free-flow speeds based on road characteristics (classification, number of lanes, land use, access control, etc.). With GPS data available from either probe vehicle studies or from GPS-based household travel surveys, planners have access to observed speed data for those roads covered in their data stream. Further, advances in data quantity and quality by private companies now allow historical travel data to be mined for speed information. Users of these data can either update the look-up tables that are used in the current models or build new tables with link- or route-specific speed values. The biggest technical challenge to either approach is building a database connection between the GPS points and the digital model network (map matching). Or, if using data purchased from a data vendor, the challenge is matching a link-based data product that uses different digital road networks. Both of these issues are discussed later in this guidance document.

In Denver, researchers used the GPS-based household travel survey data (437 households) to build free-flow and congested speed matrices by model link classification and area type (Bachman et al. 2012). Similarly, in Atlanta, researchers generated day of the week and road classification tables of congestion and speed. In both of these efforts, the GPS-based household travel survey data were used and resulted in adequate sample sizes to support aggregations of link class and area type speeds. The predominant trend with most modeling efforts is to rely on aggregate historical data from a traffic data provider due to cost, comprehensiveness, and availability. An interesting point of difference between the two approaches is that GPS from a household travel survey is based on a sample of the residents of the study area, while purchased link-based data may be a composite of fleet, commercial, and personal vehicles that traveled a segment.

Other Research or Application Areas

Good-quality GPS travel data can offer transportation planners and researchers a source of data for assessing several other travel behavior and related goals. Evaluating vehicle activity for emissions analysis or reviewing transit system operating performance can be accomplished through an analysis of GPS trace data. These examples and a long list of other possibilities require access to the data, an understanding of the data quality, and the technical capability to process the data into the summary information needed. Some selected travel-behavior–related application areas are:

- Vehicle emissions research,
- Physical activity/health research,
- Road pricing and user fees research,
- Transit ridership analysis, and
- Evaluation of traffic congestion.

For example, investigations at Georgia Tech used GPS data collected from a fleet of 480 buses from the Cobb County School District to analyze the locations, durations, and conditions of bus idling. Through the use of the MOVES emissions model, the researchers were able to quantify those emissions, as well as the emissions and fuel savings that could be had by eliminating the idling periods. The team also developed a new warning system to alert bus drivers to such events (Xu et al. 2013).

The level of environmental impact on the air produced by automobile emissions is dependent on automobile speed and traffic flow. With a large enough sample size, GPS data can be used as floating car data to establish needed speeds for analysis. A Tongji University study used data from the city of Shanghai's vast taxi fleet, 29,100 vehicles equipped with GPS devices (8% to 15% of daytime traffic flow), for their calculations. When combined with point emissions data from local sensors, the collected information was used to develop emission factors to allow the MOVES emissions simulator to account for a Chinese fleet in place of the model's American vehicular fleet assumptions (Liu, Wang, and Chen 2013).

In another instance, GPS-equipped floating cars have also been used to help diagnose traffic congestion issues. European researchers used a massive GPS data set collected by a Dutch-based location-based services company from their in-vehicle mapping software. The data include over a trillion measurements across the European Union. A methodology was proposed that quantified and made clearer the delay experienced by EU drivers during the peak period and overall. The highest levels of delay per kilometer during the 1-hour peak period were determined to be in the UK and the Benelux countries. While this study was conducted at a multinational level, the methodology could also be applied on an urban level (Christidis and Rivas 2013).

Furthermore, GPS data can be used to analyze the circumstances and frequency of unexpected delay. Analysts can investigate the GPS traces for outlier delay events and analyze their origins and locations. In such analyses, the GPS trace data may offer speed profiles of relevant corridors and comparison between similar circumstances on consecutive days.

Sources of GPS Data

Generally speaking, there are two categories of data—primary data collection (i.e., surveys) and third-party data. At this time, the options for purchasing third-party data are very limited. However, technology trends suggest that this may be a viable option at some point in the future. Given the theoretical understanding of the basic characteristics of third-party data, a table of logical tradeoffs on key factors and influencing variables is provided in Table 1. It should be noted that the logical advantages of third-party data listed in Table 1 are not weighted, and they can be severely limited by fatal flaws of data availability, accuracy, validation, and bias.

Table 1. Factors to consider when deciding to collect or purchase GPS data.

Factor	Primary Data Collection	Third-Party Data
Cost and influencing variables	**Likely Higher Cost** • Sample size (number of participants, number of devices) • Response rates • Deployment period • Languages supported • Incentive amounts	**Likely Lower Cost** • Sample size (number of data points) • Geographic coverage • Spatial and granularity • Temporal span • Variables reported
Ease of data acquisition	**More Labor Intensive** • Requires experience in project management, field data collection management, and computer programming • Requires substantial labor for deployment staff	**Less Labor Intensive** • Purchase
Ease of data processing	**More Intensive** • Intimate knowledge of when, where, and how the data were collected • Reduced uncertainty as to what processing steps need to be taken	**Less Intensive** • Unknown; substantial probing and guesswork required to determine the extent of processing that is necessary
Usability of data	**More Certain** • Biases are known • Can collect demographic characteristics of persons and households • Wider array of travel behavior data can be collected • Sample size requirements can be monitored to ensure statistical significance	**Less Certain** • Limited knowledge of when, where, and how the data were collected • Little to no knowledge of how data were treated/processed/weighted before they were delivered to the client • Biases are unknown or have to be uncovered through investigation

GPS data that can be processed and analyzed to glean personal travel behavior information can be obtained from several sources, including travel surveys and third-party data sources. Data can be collected in a systematic and statistically based method from a sample of a larger population. The most common survey to collect GPS data within the transportation planning field has been HTSs, which were described in previous sections and are the focus of much of this research. Probe vehicle studies are also used to sample operational data from transportation systems. Probe vehicle studies require researchers to develop structured sampling plans and to collect detailed data, frequently from a GPS device, along targeted routes at specific times of day. Other surveys that collect travel information from a recruited sample are health studies that use GPS devices and accelerometers to passively collect details of travel and physical activity.

Consumer product data sources include passive GPS or cell phone data collected from devices purchased by consumers, such as mobile phones and personal navigation devices (PNDs), where the intended device use was not travel survey participation or data collection but rather some other purpose (such as obtaining route guidance or traffic conditions, or social networking). These GPS (or cell phone location) data are collected by third parties, typically the device provider or communication network provider, and then processed, repurposed, and repackaged into new data products sold to other markets such as traffic information providers or transportation planners.

GPS or location data can also be collected by other means, such as in transit surveys and from transit automated vehicle location (AVL) systems. These data sources are more specialized than the first two categories and represent small portions of transportation system users (whether as individuals or fleets). Regardless, GPS and location technologies can assist researchers with targeted behavioral analysis.

The rest of these guidelines focus on the first two sources (data collected from surveys and data collected from consumer devices) given their potential coverage of broad population and their focus on personal travel. For each of these two sources, an overview will be provided on how the GPS data are collected, what data are typically available for other uses or users, and the range of costs associated with obtaining these data. A separate discussion related to privacy and confidentiality considerations with these data sources is presented in the last section of this guidance document.

GPS Data from Surveys

This section discusses the hardware, data fields, and other specifics related to GPS-based household travel surveys:

GPS Hardware
Describes hardware considerations for selecting a GPS product

GPS Data Fields
Describes the various GPS data logging attributes and fields commonly found in GPS data files

Contents of HTS GPS Files
Describes the files typically found within data sets collected as part of a household travel survey

Costs for Collecting HTS GPS Data
Describes basic costs and considerations for conducting a GPS-based household travel survey

As mentioned previously, GPS data collection methods have been introduced in surveys as a means to obtain detailed travel behavior information from study participants. The most relevant survey using this method for transportation planning purposes is the household travel survey, which is typically conducted for regional or state planning agencies to collect detailed socio-demographic and travel information for use as inputs to travel demand models. A statistically based sampling plan is implemented within these surveys to ensure that a representative sample of households is obtained that can be expanded to the larger regional or statewide population. To encourage participation and to compensate for the burden associated with household travel surveys, many travel surveys offer participants some sort of monetary incentive for completing the survey process.

There are several approaches for using GPS methods in HTSs, including using GPS data collection in tandem with traditional diary methods for a subsample of survey households to estimate underreporting in the larger sample and the use of GPS data collection to replace diary-based travel reporting. There are numerous publications detailing various aspects of these approaches, and readers are encouraged to refer to Volume I of this report for a complete synthesis of methods and references.

Regardless of the approach implemented, the basic GPS methodology is to provide each adult in the household with a wearable GPS data logger to collect second-by-second GPS data for a pre-assigned period of time (typically 3 to 7 days). In surveys that are focused more on personal vehicle travel than multimodal or nonmotorized travel, GPS data loggers that remain in each household vehicle may be provided instead of wearable data loggers. Most recently, researchers have been evaluating the use of GPS data logging apps on smartphones to replace the use of stand-alone data loggers for some participants in travel surveys.

Public agencies considering conducting an HTS with GPS methods should understand the range of approaches available as well as the advantages and disadvantages of each approach (also covered in Volume I). In addition, it would be beneficial to have a basic understanding of the desired features of GPS data loggers, the software that can be used to configure and download the data, and methods for cleaning and processing the data. Some basic guidance on these items is presented next.

GPS Hardware

There are a plethora of GPS devices on the market today, ranging from data loggers and personal navigation devices to GPS-enabled smartphones. GPS data loggers have been the most commonly used devices in HTSs over the past 15 years and have been used for many large-scale GPS-enhanced or GPS-based travel surveys over the past 6 to 8 years. Figure 2 shows three data loggers used to conduct GPS-enhanced household travel surveys in the United States during the past 4 to 5 years, including the most recent household travel surveys conducted for Atlanta, Charlotte, Cleveland, Minneapolis/St. Paul, Philadelphia, and numerous regions across Texas, as well as statewide surveys conducted in California and Massachusetts.

It is worthwhile to note that GPS data loggers without wireless communication capabilities raise fewer data security concerns for both participants and clients than do devices that can capture and transfer position information in real time (such as smartphones).

Guidelines for Identifying GPS Hardware Needs

Beyond consideration of privacy impacts of real-time tracking enabled by wireless communications, survey firms and researchers should consider the factors listed in Table 2 when evaluating GPS loggers for use in an HTS. (Note that the first three items listed are covered in

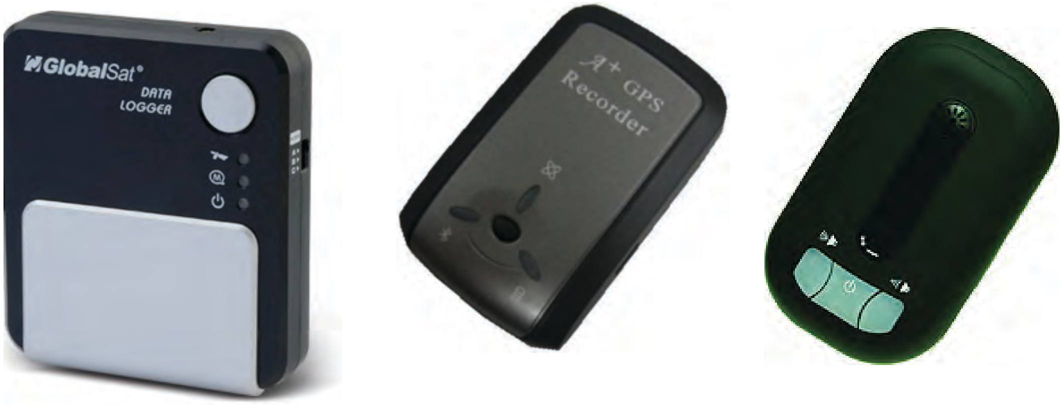


Figure 2. Example of GPS device used in recent household travel surveys.

Table 2. GPS hardware considerations.

Operational Factor	Factor Description and Guidelines for Device Attributes
Data accuracy	GPS device must be able to collect accurate locations. • More accurate data means less cleaning later on. • Desired range is from 2 m to 5 m.
Time to first fix	Time delay after satellite signal loss until the device can determine location; relevant after short or long loss of satellite signal reception (hot starts and cold starts that follow power loss or satellite line-of-sight blockage).
Positional data quality parameters	Variables such as horizontal dilution of precision (HDOP) and number of satellites can be leveraged when cleaning the data. • Always collect HDOP and number of satellites if available. - HDOP should be in the range of 0 to 4. - Number of satellites ≥3.
Battery life	The longer the battery life, the more data can be collected. • To avoid data loss, battery should support at least one full day of travel activity without the need for recharging.
Storage capacity	Memory is cheap, but redeploying devices is expensive. • Must have enough storage for the full deployment period based on logging rules and frequency. • In most cases, only records that identify movement need to be captured (i.e., select logging rules that avoid capturing stationary or zero-speed points). • The total needed storage capacity can then be estimated by multiplying the number of bytes required to store each point record by the average expected time of movement per day (in seconds) and the length of the deployment period (in days); this total can then be divided by the logging interval in seconds. • Some devices may use compression algorithms to support additional storage capacity. • Capacity to store at least 50,000 points should meet the needs of most week-long studies if zero-speed or no-travel points are not logged.
Cost	Costs are often amortized by the survey firm over several survey efforts, with lease rates calculated to cover device and accessory costs, testing, inventorying, maintenance, and losses over multiple studies. • Commercial devices range from $40 to $200, based on vendor and features supported.
Ease of and time to recharge	For multiday deployment periods, participants recharge each device provided nightly. • Recharge time should fit within this window. • Recharge should be straightforward and relatively easy to perform by participants.
Ease of use	Address such questions as: • Is it easy to tell if the device is working? • Is it easy to keep device powered on? • Is it easy to wear or carry the device?
Form factor, size, and weight	Devices should be relatively unobtrusive or else respondents may be reluctant to carry or wear them. • Review GPS data loggers available from survey specialists or online, obtain a few from different vendors or manufacturers, and try to use and wear the device as a survey participant would be instructed to do.
Ease of and time to upload data	Travel surveys often have aggressive schedules. • Ability to download the data quickly and easily is paramount.
Other features	Although these may be appealing for certain uses, each feature should be evaluated with respect to its advantages and disadvantages for use in a travel survey. • Bluetooth capability can reduce data transfer time via wireless transmission but may significantly reduce battery life for logging location data.

detail later in this document in the Assessing GPS Data Quality section.) The bulleted items can be used as guidelines when purchasing GPS devices that will be used within the context of an HTS.

GPS Data Logging Attributes

The following provides more background on some of the most important GPS data logging attributes.

Data Accuracy

Most GPS chips used in GPS data loggers, personal navigation devices, and mobile phones contain the latest GPS technology that provides location accuracy within 2 m to 5 m when a clear sky view is available. This level of accuracy is sufficient for identifying trip ends and travel characteristics, as in a travel behavior survey application, when points are logged at reasonable time intervals.

When a sky view is not available, GPS points may not be logged, or if they are logged, there may be some inaccuracy in the location calculation. This is typically caused by multipath errors that occur when GPS satellite signals bounce off buildings or other obstructions on their way to the GPS receiver. Although firmware on receivers has improved and can filter some of this noise, a minimum amount of post-processing cleaning is still necessary. Specific guidelines for assessing and cleaning GPS point data are included later in this document.

GPS Recording Interval

For surveys capturing up to 5 or 6 days of travel, a recording interval of one GPS point every 3 s to 5 s is adequate. For surveys collecting a week or more of travel, longer intervals (e.g., less frequent recording to reduce the amount of data written to the device) may be necessary due to storage capacity constraints; however, lower data frequencies may affect the performance of automated travel mode identification methods. Further consideration should be given to recording intervals required for secondary uses of GPS data, including calculations of acceleration or for drive train analysis.

GPS Device Configuration and Download Software

Many of the features included in software provided with off-the-shelf GPS data loggers may not be useful to transportation researchers (e.g., photo tagging, Google maps interface), and some of the default features (such as automatic track assignments) may make using the software cumbersome or challenging. Other features that would be useful to researchers may not be accessible with out-of-the-box software. Therefore, it is important not only to test the devices themselves but also to test the software provided with the devices for configuration options and download parameters. Often, the development of custom configuration and download tools is very helpful, if not necessary, for large-scale survey efforts. However, creating such tools can be technically difficult and can only happen if the manufacturer provides an application programming interface or documentation on the protocols used for communicating with, controlling, and transferring data from their devices. (Most GPS loggers use serial communications over USBs.) If researchers want to use the manufacturer-provided off-the-shelf software, they should be certain to choose a device that allows controlling the logging frequency, speed screen rules, and recorded variables and units.

Importing and Processing GPS Data

Loading GPS data into a project database is dependent on the software used to download the data, or more specifically the data format output from the software. Ideally, the software will allow for the data to be downloaded in multiple formats, including comma-separated values or text files that are standardized so the data can be quickly reviewed for compliance and completeness and that put the date/time and coordinate information in a standardized format. When this is the case, the data can then be quickly imported into a database for viewing in a number of different software packages. Once the raw data are downloaded and imported into a project database, data processing is necessary to compile point data into trips that can be used as needed in the overall survey design. Procedures for cleaning and processing GPS point data, for identifying trip ends, and for imputing missing trip details and socio-demographic characteristics (when not available) are provided later in this document.

Contents of HTS GPS Data Files

GPS data collected in household travel surveys typically start as raw GPS trace data with a row for each point captured. Most GPS data loggers pull only the essential variables that are output from GPS receivers: date, time, latitude, longitude, heading, speed, altitude, horizontal dilution of precision (HDOP, an accuracy measure of GPS satellite geometry), and the number of GPS satellites in view.

These point files (each file containing all GPS points collected by a GPS-instrumented survey participant or vehicle) are typically imported into the project database and then processed into trips and trip details. If the GPS component design is to compare these GPS trips with travel-log–based reported trips, then an attempt will be made to match these GPS and log-based trips (most likely through a combination of automated and manual methods), with the results saved for use in creating trip rate correction factors.

How GPS data are delivered is largely client dependent. Many clients prefer to have the processed GPS data as well as some level of raw GPS data delivered. A fully processed GPS data deliverable is most commonly provided in an Access database or a statistical software format such as an SAS file and consists of GPS-based trip details, trip segment details, and point data associated with the trips and trip segments. Identifiers that associate the trip to a household and person, as well as travel mode, are typically included as part of the deliverable. A raw GPS deliverable can include delivery of the GPS data as they were when downloaded, without having filtered any data points that did not represent real travel (noise and poor quality points or points recorded while the participant was at a destination) or that took place outside the prescribed travel date range. Both processed and raw GPS data can be viewed in many software applications, such as Google Maps and Google Earth, a commercial geographic information system (GIS) (e.g., the suites from ESRI, Autodesk, and Bentley), open-source GIS tools (e.g., Quantum GIS, gvSIG, Open Jump), and several statistical packages (e.g., SAS, R, SPSS).

Costs of GPS-Based Household Travel Survey Data Sets

The costs of obtaining an HTS data set do not have to be tied to actually conducting an HTS. However, most practitioners interested in the potential of using GPS data to replace travel surveys think of these two items as one and the same. Therefore, this section will start with an overview of the cost factors in a household travel survey. The costs of adding a GPS component to a traditional diary HTS or conducting an HTS as a GPS-based survey are highly dependent on a range of HTS design parameters. The following list provides a summary of the main HTS design parameters that will affect the cost of implementing a GPS-based component.

- Sample sizes of the pilot and main survey.
- Length of the field data collection period for pilot and main survey.
- Languages to be supported.
- Anticipated response rates.
- Incentive amounts.
- Number of days of GPS data collection per household (deployment period).

Due to the number of variables as well as the interdependencies between variables that affect cost, firms that specialize in these types of surveys are reluctant to provide price quotes in the absence of detailed survey specifications, especially in the cost-per-household rate that most HTS clients prefer. Furthermore (and understandably), there is a reluctance among the survey research firms to share detailed cost information due to the highly competitive market space.

Nevertheless, based on available information for travel surveys conducted over the past decade, a traditional diary-based survey has costs in the range of $150 to $225 per household, with GPS add-ons (which involve adding GPS devices to diary-based methods) adding $200 to $350 per GPS household to the total cost. GPS-based surveys that do not require diary recording and reporting range from $300 to $500 per household. (There have only been two GPS-only surveys conducted in the United States to date, and the costs mentioned here do not factor in the sample size expansion potential when collecting multiday GPS data, which may prove to reduce the cost per equivalent household travel day.) It is suggested that agencies considering these types of surveys contact travel survey research firms directly to obtain price quotes based on their specific survey needs.

It is worth noting that researchers have several options for accessing GPS data sets once a GPS-enhanced or GPS-based HTS is complete. One option is to collaborate with a sponsoring agency of an HTS, with data confidentiality agreements exercised to protect participant confidentiality. Another option is to contact the Transportation Secure Data Center (TSDC). Several years ago, the FHWA collaborated with the National Renewable Energy Laboratory to provide a centralized, secure place to store GPS data sets collected within transportation studies. Consequently, the TSDC was established to allow researchers to analyze GPS data collected in travel surveys in a secure environment that protects the privacy of the individual study participants. Interested researchers are required to complete an application explaining the intended research and use of GPS data; once the application is approved, there are no costs for accessing and analyzing these data sets. Not every GPS data set collected in an HTS has been submitted to the TSDC, although efforts are well underway to increase its inventory. More information about this center can be found at http://www.nrel.gov/tsdc.

Third-Party Travel Data

Third-party travel data are data that have been archived by a private company and are available for purchase. These data are generated from historical traces or real-time feeds of travel information generated by consumer products such as cell phones, smartphone apps, and telematics devices. These sources are becoming more widely available to travel behavior researchers, and this section provides some guidelines and instruction for those considering a third-party data source.

Link-Based Data Products	Describes data that can be purchased for roads or road segments in a study area
Point-Based Data Products	Describes GPS data that can be purchased from private companies
Cellular-Phone-Generated Data Products	Describes cellular-based data products
Third-Party Data File Contents	Describes the common contents of data files from private companies

Over the past decade, consumers have purchased personal devices such as smartphones and personal navigation devices at an astonishing rate. Each of these devices is capable of capturing detailed location, time, and speed information at regular and frequent intervals. Providers of these devices (or of the wireless communication networks that support real-time information access or exchange with these devices) have capitalized on the massive quantity of location data available from these consumer bases. Consequently, private firms have collected these crowdsourced data, processed the data into new data products (essentially repurposing the data), and have been marketing these products to industries such as traffic information and transportation planning.

Unlike GPS data collected for travel surveys, which involve active data collection efforts and are not purchased from a private entity, link-based data products, processed GPS data, and cell-phone–based location data are available for purchase from a variety of private companies.

The specific products available at any given point in time are variable and seemingly increase in their capability for supporting personal travel behavior research. However, the value of these data for research is mixed due to the fact that there are market-driven and public-policy–driven needs to protect the privacy of the original data source (i.e., individual users of phones or PNDs). Companies that sell data like this must protect the privacy of their raw data sources to ensure the data's future availability. This need for privacy protection runs counter to the needs of travel behavior research, which prefers full detailed traces on where and when these individuals travel. Despite this limiting factor, there are products available that are targeted toward transportation planners. New or improved products will likely be available over time.

Given that these data products come from private companies that do not share processing details and are constantly changing or improving their data offerings, this guidance cannot discuss specific products or detailed processing techniques for use with specific products. Instead, this discussion categorizes known product categories and discusses key elements of interest that define their usefulness. The processing and planning application techniques described in previous sections may not apply to consumer-data–based solutions since the quality and characteristics of these data are highly variable. Users of these data sources will have to apply good judgment and use this guidance as a starting point for a decision and plan to purchase and use data from a private company.

Link-Based Data Products

Link-based data products are databases that can be purchased and for which raw information has been aggregated to road segments (as opposed to GPS points or sensor locations). Almost all link-based data products that are currently available originate from efforts by private companies to build real-time traffic solutions. The real-time traffic market promised significant revenue for those companies that could deliver reliable data in real time or near real time. As part of the original planning and growth of real-time traffic solutions, communication and location-referencing standards [i.e., radio data system–traffic message channel (RDS-TMC), Alert-C] were created and universally accepted by data gatherers, providers, and broadcasters. These standards dictate how speed and other traffic information are spatially referenced and communicated in low-bandwidth solutions consistent with in-vehicle data consumption scenarios. These standards do not conflict with potential application for transportation planning but result in data definitions and terminology that are not standard in the transportation planning community. Further, procedures used for processing raw GPS data from fleet vehicles and personal vehicles in these products were targeted specifically for the real-time traffic market.

The resulting data products, available for purchase by transportation planners and travel behavior researchers, are link-based speeds reported at some time interval (typically 2 min to 5 min). Over time, these same traffic data were aggregated into historical data products with typical transportation planning time periods (e.g., day of week, hour of day, season) and marketed to transportation planning professionals to support congestion management programs and model development efforts. It should be noted that the FHWA's National Performance Management Research Data Set is a link-based data set developed from a third-party data set and available to departments of transportation and metropolitan planning organizations. This data set was purchased from Nokia and includes 5-min speed data for TMC segments along all National Highway System roads. While the data have limited travel behavior research capability, they represent a typical link-based data product.

While there has been some effort to change the location-referencing schemes that rely on the RDS-TMC and Alert-C standards, most link-based products use the TMC reference. In this standard, road segments are identified with a TMC code that typically is made up of four concatenated codes to create a unique ID. The four codes are a one-digit country code, a two-digit

TMC table code (region identifier), a one-digit direction code (indicating positive or negative to identify the correct side of the road), and a five-digit location code (a single point on the road network that is unique for each table). Users of this data will have to translate this location-referencing scheme to their own base road network. Most data providers will assist with this task, and a GIS file with these references can be acquired.

Link-based data products originate from a wide variety of sensors and GPS data sources that can include commercial fleets, taxis, buses, personal vehicles, and delivery vehicles. Early products relied heavily on data from commercial vehicle fleets. In more recent times, data from personal vehicles have been available to these companies and added to the mix of raw data sources. It should also be noted that most companies that generate real-time traffic speed and delay estimates also generate estimates of traffic speeds when raw data are not available. This estimation can be reflected in the aggregated data products and may result in reduced variability metrics.

Given these issues, planners and researchers considering the purchase of link-based travel data are encouraged to:

- Be sure that the fleet makeup of link-based speeds matches their needs.
- Understand that some values may have estimated speed or travel time values as opposed to data generated from actual field observation. Some data products will have identifying information for when these conditions occur.
- Be sure that the raw links (road segments) can be translated into definitions that match internal road definitions. This can be a time-consuming issue if direct translation is not available.
- Be sure that they have some level of understanding regarding the raw sample sizes used to generate the data. Lower-volume roads could have lower sample sizes if the original raw data sources rely on fleet GPS data sources.

GPS Point-Based Data Products

Several companies offer raw point-based GPS data products that are arguably more useful than link-based products for personal travel behavior analysis. The sources of these raw trace data are GPS-equipped devices (personal navigation devices, smartphones running embedded applications, embedded GPS devices in vehicles, etc.). There is strong potential for these data products to describe detailed travel behavior (origin–destination, path, trip characteristics, etc.). However, the challenge of protecting the privacy of the raw data sources remains a key complication in the release of useful point-based data products. The market value of providing full-trip or device-based travel details combined with liberal legislation on the use of consumer data may one day ease access. Currently, companies rely on broad license agreements from navigation devices and navigation applications in smartphones to protect and justify the data archiving and repackaging for distribution. However, a sense of caution still permeates the data market as one or two high-profile privacy concern cases can significantly affect public opinion and result in data resell restrictions. This has already happened to some degree with Google and Apple.

There are several techniques used by companies to prevent users from piecing together a full trip trace for a participant. One approach is to provide the raw GPS data in somewhat large time intervals or frequencies (e.g., a point every 5 or 10 min). Device/user IDs are also altered for each interval, making it difficult if not impossible to piece together an individual trip trace. The data are segmented or tagged in this way because these snippets of data can still be used to support the primary markets of real-time traffic.

Additionally, raw data may be cleaned by the provider by removing data points from the first few and last few minutes of a given trip. This prevents users from identifying trip ends and only allows the on-road travel portion of the data to be available for analysis. A third approach for privacy protection of point data is to eliminate the person or device ID entirely so that all that is available are coordinates and other GPS attributes (e.g., timestamp, speed, heading). This approach prevents someone from extracting any trip details and forces data users to rely on the individual point values (such as speed) for their needs.

Given these issues, transportation planners and researchers are encouraged to:

- Understand the fleet makeup of the vehicles from which the data are being captured (personal vehicles versus commercial vehicles).
- Understand any perturbation techniques that have been applied to protect privacy and be sure that these techniques do not prohibit the intended use of the data.
- Identify the original technology source for the point data (navigation device, cell phone signal, GPS-based application, etc.) and understand the accuracy implications.
- Understand the privacy protection policies of their organizations to ensure that the purchase of point-based data from private citizens is acceptable.

Cell Phone Data Products

Location data generated by cellular phones are being marketed as a valuable tool to transportation planners for understanding the movements of individuals around their environment. These non-GPS data are generated when cellular phones are powered on and are being actively used (such as in use for a phone call, for messaging or web browsing, or when running an application). In general, a position, device ID, and timestamp are created through triangulation from cell towers during these activities. A moving device that is active will generate a trace of positions over time. Relative position can also be generated when a moving cell phone is handed off between towers. By reviewing a cell phone's complete historical trace, a trace of an individual's movement can be estimated. An attractive element of this data source is the ability to generate traces for very large segments of the population. While the second-by-second location of a device is not typically captured, analysis of device activity over weeks or months can reveal movement patterns. This capability has the potential to assist transportation planners by capturing zone-to-zone interchanges for those individuals with mobile phones managed by participating providers.

As with the other data products, privacy protection is a major concern for the companies gathering, processing, and selling these cellular phone data. For this reason, individual traces are not currently released. Instead, data are aggregated into zones, such as census block groups or traffic analysis zones, and zone-to-zone trip totals and travel times are estimated. Current products can also break the zone-to-zone trips totals into specific time periods and estimated trip types (e.g., home-based work, home-based school). Most of the processing procedures for the various data aggregations and disaggregations are based on proprietary data mining techniques, and independent validation studies have not been completed. Several research studies have been completed with positive, though cautionary, reviews (Asakura and Hato 2004; Ratti et al. 2006; Gur et al. 2009; Ma et al. 2012). The data that can be purchased have been processed using private and patented algorithms that are not likely to be released. Therefore, users of these data should plan to conduct independent validation steps or approaches.

The following issues should be considered for users of mobile device or cell phone data:

- Understand the market penetration of the user base in the study area (i.e., which phone carriers are represented in the data and the proportion and characteristics of the population covered by them).

- Understand the expected sample size given the number of zones, time periods, and trip purposes desired.
- Identify the end-solution impact of having longer periods of aggregation (weeks or months) as opposed to more traditional, shorter time-period surveys.
- Develop an independent validation framework to identify underrepresented or overrepresented demographic segments. Validation using U.S. Census data, previous HTS datasets, and model outputs may be helpful in identifying potential biases in data.
- Understand the fleet makeup of the data (personal travel or from freight/commercial vehicle movement).
- Understand the cell phone carrier coverage area since entering or departing a coverage area may cause difficulty in capturing the full trip.

Contents of GPS Data Files from Private Firms

Commercial GPS data file formats vary with each company and each product offering. Very few standards exist since most of these products are custom developed for each client and change based on improvements in technology, market demand, and raw data availability. For products that originate from aggregated solutions from real-time traffic data services, the native format and content are likely to use an XML schema that contains the following data elements:

- TMC location reference (usually a location code, not coordinates)
- Timestamp in coordinated universal time (UTC)
- Speed or travel time value
- Congestion index
- Event code

Contents of Aggregate Data Files

Note that aggregate data do not allow a user to build a trace of data points or travel segments for a single user. Aggregated data of this sort do not typically meet the needs of planners and researchers for detailed or individual travel behavior analysis except when used to build baseline conditions for travel demand models, to validate models, or when used to evaluate time-of-day, day-of-week, and seasonal variation.

More detailed aggregate data were used in SHRP 2 Project L04, "Incorporating Reliability Performance Measures in Operations and Planning Modeling Tools," conducted by Delcan Corporation. Delcan purchased TomTom data trajectories where each record represented a trip segment with a unique ID that corresponded to a road segment from a GIS database. The data set was provided to Delcan under very strict usage agreements to protect the privacy and external release of the data. Similarly, aggregated cell phone data can come in a number of formats and can be customized for each client. In general, though, these data are typically aggregated into counts of zonal trips as:

- Origin zone ID,
- Destination zone ID,
- Time period,
- Trip type, and
- Number of trips.

Contents of Disaggregated Data Files

Disaggregate products, as mentioned previously, are not easily accessible due to privacy concerns. Often, such disaggregate data may be shared or sold between private firms that are using the data for other data products. The products that fall in this category may come in a variety of standard text file (e.g., fixed width or delimited) or binary database formats (e.g., Dbase, Access). The content of these files, as expected, may vary depending on the original source. These are typically generated from raw data collected by GPS-based devices where the original data have been manipulated or truncated in some manner to protect the privacy of the original source. The basic elements of these data will likely include:

- Device ID,
- Longitude,
- Latitude,
- Timestamp,
- Speed, and
- TMC (possibly).

As mentioned, these files have likely been adjusted so that the device IDs are reset on a regular basis (e.g., 10-min intervals or once a day) or based on spatial aspects (e.g., local roads or off-network points excluded). It should be noted that most companies providing consumer-generated data have many options for end-product data structures. Some companies offer fully processed interactive analytics services covering data for links, corridors, and regions. These products have varying data structure options as well as a variety of online tools for extracting key metrics.

Costs of GPS Data Sets Obtained from Private Firms

Private companies that are selling or reselling GPS data products have restrictive data licensing agreements with pricing that is market based. The determination of market value for aggregate consumer-generated travel behavior or traffic data has been hard for private companies to define since they try to compete with decades-long standard survey research and modeling practices and are also forced to use public-sector project funding schemes, which can be inflexible. Furthermore, there are theoretical challenges to new approaches and suspicions about the processing steps used to build these products. These hurdles are being overcome as independent research projects are gradually helping to assess the value and supported uses of consumer-generated products within the field of transportation planning and research.

Data costs from private firms are highly variable but seem to be related to the size of the data request. Companies can adjust pricing of their origin–destination data based on the level of aggregation requested and the population of the region. More aggregate data sets for smaller populations are less expensive than low levels of aggregation for larger populations.

As mentioned previously, products and pricing change frequently. However, users should expect regional data costs to be between $50,000 and $500,000 for aggregate origin–destination data that cover a medium-to-large metropolitan region and are suitable for supporting travel demand model needs. Other products are priced either by the route or for a full region. The following statements regarding pricing of products were provided by consumer data product resellers as part of a 2012 survey conducted in support of this NCHRP project:

- **AirSage origin–destination data product:** "Cost structure varies from project to project and is based on multiple variables, such as population, study area, study length, etc."
- **INRIX historical analytics suite:** Available as a subscription service.

- **Nokia historic traffic data:** "Licensing fees are determined based on the number of users, licensing term, geographic extent, and data delivery mechanism."
- **TomTom custom travel times:** Products are priced on a per-route basis.
- **TomTom custom area analysis:** Products are priced based on geographic scope, road class, and number of queries.
- **TrafficCast Dynaflow historical data:** Product is priced based on road miles and data update frequency.

Given that this list of companies is not exhaustive (and is ever changing) and that product offerings and pricing can vary significantly over time, interested users should always contact the companies directly for product descriptions and specific pricing based on the user's needs.

Assessing GPS Data Quality

While GPS data points can be exact, spatially and temporally speaking, certain environmental conditions can affect data accuracy. These conditions are most common in areas with no direct satellite line of sight. Tunnels will prevent the collection of GPS data altogether. Urban canyons cause distortionary multipath effects brought about by the presence of tall objects that reflect or block GPS signals, causing them to arrive at the antenna with varying levels of delay. Examples are tall buildings, tree canopies, narrow streets, and actual earthen canyons (Quddus, Noland, and Ochieng 2006). This section reviews indicators and techniques for evaluating and cleaning raw GPS data.

Quality Indicators for GPS Data	Discusses methods for evaluating raw GPS data quality
Cleaning Raw GPS Data	Discusses methods for cleaning raw GPS data

Quality Indicators for GPS Point Data

The quality of location or travel data collected using GPS technology has a direct impact on the results and effectiveness of analyses based on these data. Some aspects of data quality can be assessed at the individual point level, while others require that points coming from the same rover (i.e., GPS-instrumented person or vehicle) be examined in context.

Measures of Spatial Accuracy

The quality of GPS point data is usually quantified in terms of the spatial accuracy of the reported coordinates. If available, positional quality indicators such as horizontal dilution of precision and number of satellites should be used. The HDOP value of a computed point indicates how well the satellites used were dispersed in the sky. The further apart and well distributed they are, the lower the number will get; conversely, situations where all satellites used are close together or perfectly lined up will result in large numbers.

With HDOP, the smaller the numbers the better the quality of the reported coordinates. On the other hand, a higher number of satellites used to compute a position usually indicates less error. A minimum of four satellites is needed to compute a three-dimensional coordinate. This is because the receiver also has to solve for the time error in its internal clock; however, if the receiver only computes a two-dimensional coordinate (holding the last known height constant), it can compute a point solution using only three satellites.

When selecting thresholds for HDOP and number of satellites, one must take into account the type of application and where along the collected paths each point is. For traditional travel survey applications, points that are used to determine the origin and destination should be of better quality than those collected along the traveled path. However, path points are equally important if the goals of the project include the analysis of route choice. A conservative starting point for these thresholds is the values proposed by Stopher, Jiang, and Fitzgerald (2005) of HDOP values lower than three and a minimum number of satellites of five; these values can be used when a high degree of accuracy is desired. However, it is still possible to obtain reasonable quality traces when allowing points with HDOP between zero and five and a minimum of three GPS satellites.

When examined in context, the quality of GPS points can be rated in terms of the amount of noise in the captured path and the presence of unrealistic movement. The presence of points with poor, or varying, positional quality in a trace can be visualized by connecting the points that make up a trip to an activity location with a series of line segments. Once this is done, the error in the individual points can be observed in the form of waviness and roughness in the represented trajectories. Similarly, positional outliers manifest themselves as spikes in the plotted path, where the trajectory quickly deviates from the prevailing path and just as quickly returns. Figure 3 depicts the noise inherent in the GPS data collected during a given trip; for this trip, the level of waviness is quite reasonable, and the likelihood of matching this trace to the underlying transportation system network links is high.

The first type of issue is usually of little concern and does not tend to add much error to the captured trip distance and consequently derived average speed estimate. However, instances of single point outliers in a captured trajectory can add significant error to the accumulated distances and should be removed from consideration. This can be done by comparing GPS receiver point speeds with speeds estimated by dividing the distance between subsequent points and the time interval between them; instances where the distance-derived speeds are much higher than the reported speeds should be investigated. The data cleaning method proposed by Lawson,

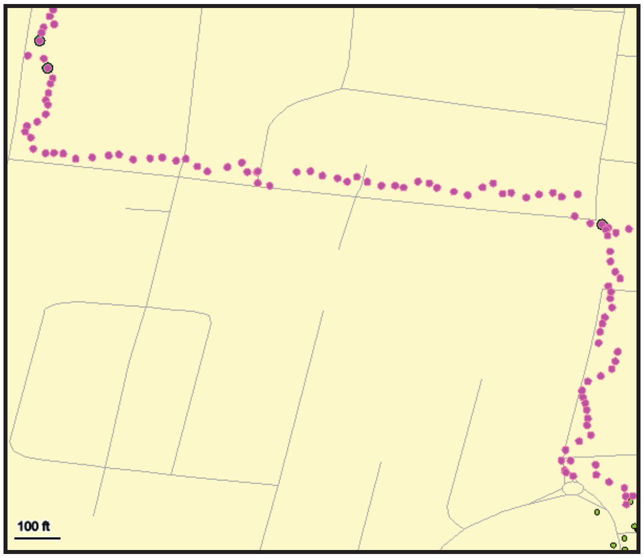


Figure 3. Example of wavy GPS trace.

Table 3. Spatial accuracy indicators for GPS point data.

Indicator	Description
Horizontal dilution of precision	• Measure of the level of satellite dispersion throughout the sky. – Lower values reflect enhanced spatial accuracy.
Number of satellites	• Higher values correspond to enhanced spatial accuracy. – Two-dimensional coordinates require at least three satellites in view. – Three-dimensional coordinates require at least four satellites in view.
Noise	• Presence of points with poor or varying positional quality. – Noise is manifested as roughness or waviness in path of GPS traces.
Unrealistic movements	• Sudden spike away from path followed by a quick return back to path. – These can add significant error to accumulated distances and average speed estimates.

Chen, and Gong (2010), and evaluated in Volume I, can be applied to remove these points from the input stream. Table 3 presents a summary of spatial accuracy indicators that are applicable for GPS point data.

Coverage

Coverage of a GPS point data set can be defined in spatial or temporal terms and is important for verifying that the data covers the expected times and locations needed for the data's intended use. Determining the coverage of a GPS data set can be done using commercial or open-source GISs. These tools allow users to spatially visualize the points in the data set and relate with other existing spatial data sets. Spatial coverage can then be expressed in terms of the limits of the area that the data set covers, either in the form of a polygon or of a bounding box (i.e., the rectangle formed by the lower left and upper right coordinates of the area). The temporal coverage consists of the time bounds of a data set. The first concept is simple to quantify and can be done by sorting the points by data and time and recoding the date and time of the first and last points in the set. The coverage found this way can then be reported in terms of the data's start and end timestamps.

These two definitions of coverage can be used at the aggregate level (i.e., operating across point data collected by multiple GPS devices) and at the individual trace level. One may also define coverage in terms of the completeness of the travel patterns captured by individual traces. This can be done by determining the presence of temporal and spatial gaps in the data stream so that trips and tours derived from it will ultimately produce complete tours (i.e., data for a particular day starts and ends at the same place, and traces exist leading to and from all identified activity locations and times). Coverage in these terms can be estimated by dividing the size of the spatial gaps in the stream by the total captured distance; this can be estimated by travel day or across a given GPS device's points.

Measurement Error and Biases

The main sources of measurement errors in GPS data are related to less-than-ideal satellite reception during data collection. The main factors that affect reception are antenna position and environmental conditions. Recent advances in antenna technology have made positioning less of an issue, especially with small and wearable GPS devices. This means that the environment is the main source of measurement errors in GPS. Examples of environmental conditions that introduce measurement errors are dense urban canyons, heavy tree foliage, and complete sky blockages such as those caused by tunnels and overhead structures. Urban canyons cause an error condition referred to as multipath, under which satellite signals are reflected off hard structures and arrive at the antenna with varying degrees of delay, and can introduce significant positional

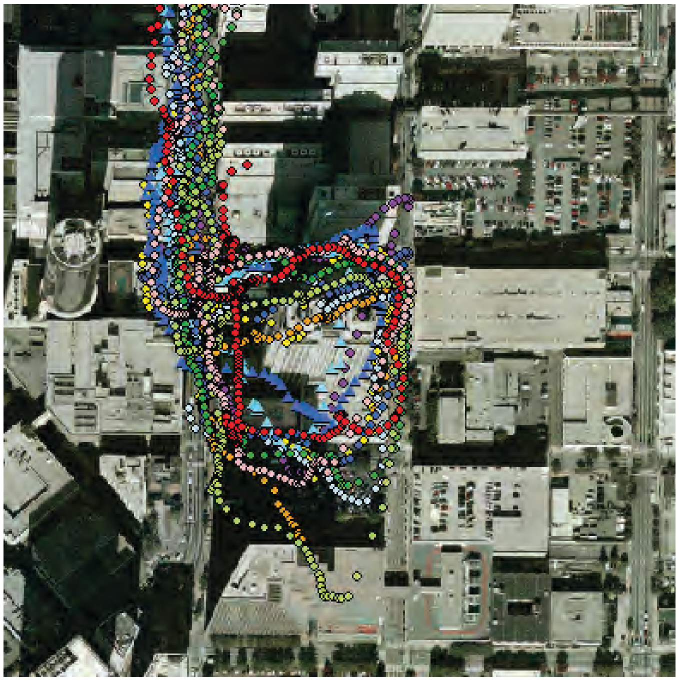


Figure 4. GPS traces collected from multiple receivers in the same vehicle.

estimation errors. In addition, urban canyons also limit the area of the sky visible to the antenna, further restricting the ability of the receiver to generate a good navigational solution.

Figure 4 shows the effects of an urban canyon on collected GPS traces from the testing of multiple receivers in downtown Atlanta in 2008. This test was performed by GeoStats to evaluate the performance of multiple GPS loggers in the context of person-based data collection for HTS and physical activity studies. Individual traces can be identified by point shade and symbol, and they clearly show how much path deviation urban canyon effects can introduce into the point solutions of different devices under the same data collection conditions.

In addition to two-dimensional coordinates, GPS receivers typically output speed and altitude. The first is usually computed internally by the receiver using the Doppler effect on the satellites' received signals. Due to the way in which the speed is internally computed, it tends to be much more accurate than speed values derived by comparing subsequent points and timestamps (Zito, D'Este, and Taylor 1995). As a general rule of thumb, one should expect the accuracy of speeds and positions to increase with higher reported speed values. This is because higher speeds are usually associated with open spaces, which are ideal for GPS reception. Exceptions to this can occur if a high-speed vehicle is moving in an out-of-sky-view area such as a tunnel or urban canyon.

The accuracy of altitude values reported by GPS receivers, on the other hand, cannot be estimated independently. This is because the GPS system estimates altitude as height above the ellipsoid of reference, which is a smooth mathematical approximation for the surface of the Earth. However, true height can only be determined as the difference in level between the surface of the Earth, referred to as the geoid, and the ellipsoid. This makes the accuracy of GPS-derived altitude dependent on the accuracy of geoid elevation data in the area where the data were collected, and hence, altitude is harder to assess without further research and data gathering. Finally, even though a minimum of four satellites is needed to estimate a three-dimensional coordinate (latitude, longitude, and height), most receivers will still output point solutions when only three satellites are in view while imputing height (e.g., holding the last known value constant). As a consequence, height values from points where only three satellites were used to estimate the solution should be discarded.

Although data collected using GPS technology are, by nature, direct and free of biases, bias can be introduced in a GPS data set through poor sampling of households and their vehicles or persons. A known issue is that sampling senior households and households with children can be difficult. This finding has led to study designs where older adults and children have their travel patterns collected using traditional HTS techniques (and proxy reporting in the case of children), while other adults and teenagers use GPS devices. Sampling biases also exist in automatically collected trace data from vehicle and cell phone tracking services, so it is important to obtain as much information about them as possible and to use weighting in any subsequent analyses that may be done using these data.

The following guidelines and considerations should be taken into account when using GPS data:

- Local environmental conditions may hinder data collection.
- Urban canyons can also distort GPS absolute and relative positional accuracy.
- Tunnels and hardscape sky blockage will likely prevent GPS point collection.
- Dense tree cover may distort the GPS absolute and relative positional accuracy.
- Speeds computed internally by the GPS device should be trusted more than those based on subsequent points and timestamps.
- Discard height values derived from points with only three satellites in view.

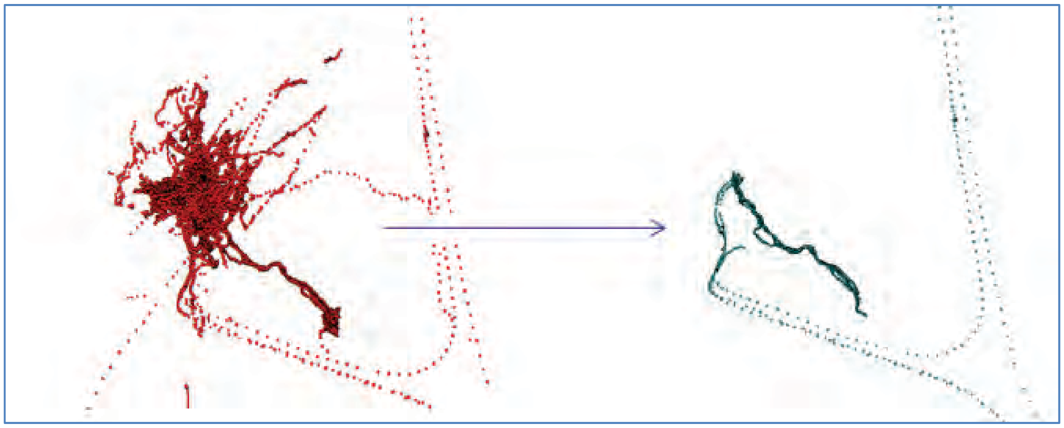
Cleaning Raw GPS Data

The first step involved in preparing raw GPS data for analysis is often referred to as data cleaning. The main goals of this step are to convert the data into a format that is convenient for processing and to remove points that do not represent real movement of the receiver or points that are of poor positional quality. A simple approach is to compare consecutive points and test for realism. Speeds that are significantly greater than the local speed limits and accelerations that are incompatible with ground transportation modes are indicators of errant positioning. Additionally, sharp turns at high speeds are unlikely to be accurate representations of a participant's path and can be deleted. Likewise, points containing trivial information can be singled out and eliminated from a data set. Points on a straight line at constant speed can be removed for investigations into travel routes. Points at intersections and destinations are more relevant to that type of analysis and should be sufficient to extract the necessary information.

The tasks involved in achieving the first goal may vary depending on the original data source being used, but in general terms it will likely involve some of the following steps:

- Convert UTC time to local time using time zone and daylight savings information for the area where each point was collected while preserving original UTC time values.
- Convert speed values to desired analysis units (some units report speed in knots).
- Calculate time interval between points in seconds.
- Compute distance between points in desired distance units.
- Estimate point accelerations using subsequent points' speeds and time intervals.
- Remove points outside the prescribed data collection date range (i.e., exclude points that may have been captured as the device was in transit to and from the participating household).

The tests performed as part of this research project indicate that simple rule-based methods, which leverage GPS data-quality metrics such as HDOP and number of satellites, as demonstrated



Source: Abt SRBI presentation for Delaware Valley Regional Planning Commission.

Figure 5. GPS points before and after data cleaning process.

by Lawson, Chen, and Gong (2010), were the most effective. For cases where HDOP and number of satellites information is not available, the method proposed by Stopher, Jiang, and Fitzgerald (2005) can be used since it achieved good results using only speed, longitude, and latitude. Given how effective HDOP and number of satellites are in helping to filter out poor quality data, they should be collected as part of the logged GPS point stream whenever possible.

In addition to filtering out poor quality points, this step can include the removal of points with low speed (e.g., lower than 1 mph is a commonly adopted threshold), as well as the computation of local time from the UTC timestamps captured by the GPS receiver. This last step is particularly important to latter processing steps such as trip end and trip purpose identification.

Figure 5 displays the results of the data cleaning process. The points on the left represent the full GPS trace that was collected before, during, and after trips made to a given location. The points in the image on the right represent the GPS trace remaining once the cleaning process removed all detected noise.

Processing GPS Data

This section provides guidance regarding the processing of GPS data to extract key travel behavior information important to transportation planners and researchers.

Inferring Travel Attributes	Describes methods for extracting travel attributes from cleaned GPS data
Inferring Tours and Activity-Travel Patterns	Discusses identifying travel patterns in processed GPS data
Inferring Demographics	Discusses estimating demographic information from raw GPS data
Map Matching	Describes map matching and linking raw GPS data to road networks

GPS data processing involves utilizing cleaned GPS travel data to infer useful travel characteristics of individuals or households. There are many options for analysts to process GPS data, and few have been fully documented or standardized. The guidance provided in this document is organized around the three primary processing steps (detecting trips and inferring travel attributes, inferring tours and activity-travel patterns, and inferring demographics) required to convert a raw GPS data set into a travel behavior data set. The fourth topic covered in this section is how to match raw trace data to road networks to obtain network link and route information. Although not necessary for household travel surveys, map-matching results are quite useful for route choice analyses as well as for calculating network travel time and performance metrics.

The methods described in this section target leveraging or generating information that would be obtained through a traditional household travel or activity-diary survey. While these methods can be used for data collected from other sources, the overall sequences provided and underlying assumptions are for complete household travel data. Furthermore, the guidance within this section is primarily aimed at the more complex task of processing person-based GPS points (as opposed to vehicle-based GPS data) since this approach is the most challenging and the most likely format for future GPS data products.

If it were possible to accurately infer all three types of travel behavior information presented in this section (i.e., trips, tours, and demographics), the effect would be to replicate a household travel survey using anonymously collected data and publicly available extant data. If fully and properly implemented, these methods have the potential to supplement or replace the traditional household travel survey, greatly reducing the burden of collecting such data and potentially increasing the sample sizes available.

The research and test results from this project indicate that the potential does exist for GPS data processing methods to one day achieve the goal of generating a full supplemental or replacement travel behavior data set. However, the methods and guidance provided in this document cannot achieve this goal given that the state-of-the-art in this topic area is not quite there, especially for the more challenging tasks such as trip purpose imputation and inferring demographics, and readers should be aware of this. Furthermore, it should be noted that this research is based on travel behavior at the person level (i.e., based on individual GPS traces), not at the household level upon which household travel surveys and travel demand models are based.

For more details on the methods and results of the exact tests performed for this project, readers are encouraged to refer to Volume I, Chapter 3, of this report. In addition, instructions for the scripts and code created for these tests and provided as part of this project are covered in Volume I, Appendix F.

Prerequisites for Successful GPS Processing

The ability to successfully implement each of the three primary processing steps mentioned previously is highly dependent on the type of GPS data collected, with each process imposing increasing requirements on the GPS traces. The inference of individual trip attributes imposes the minimum level of restrictions on GPS data. The procedures discussed here can be applied to any set of GPS traces, or even to non-GPS traces (i.e., cellular traces), provided that the traces represent complete trips or sets of trips and are not based on random polling of devices. The ability to convert inferred trips into tours and daily activity patterns, however, is restricted by the need for the GPS traces to be linked via a unique identifier (which can still be anonymous) or to come from continuous tracking of the same device. This precludes the use of some data sets that do not link trips uniquely. Finally, the ability to associate demographic information with a trip using the methods presented in this document requires a minimum of 1 day (and preferably more than 1 day) of fully linked GPS traces and the ability to identify home locations because these methods rely on matching daily activity pattern characteristics.

Fusing GPS Data with Other Databases

The ability to enrich the GPS data with links to other sources of information and databases greatly enhances the methods discussed in this report. The spatial nature of GPS data provides a natural link to many other information sources that contain a spatial component, such as census data, land use information from planning agencies, TIGER Line files, and many others. These other data sources form the core inputs to the methodologies described in the following sections. Geographic information systems and spatially enabled databases make the process of joining these ancillary data with GPS traces fairly straightforward. In addition to their spatial nature, GPS traces also feature a temporal dimension that can be used to link trace data to time-dependent data sources such as transportation performance data sets (i.e., traffic and transit performance data) and weather.

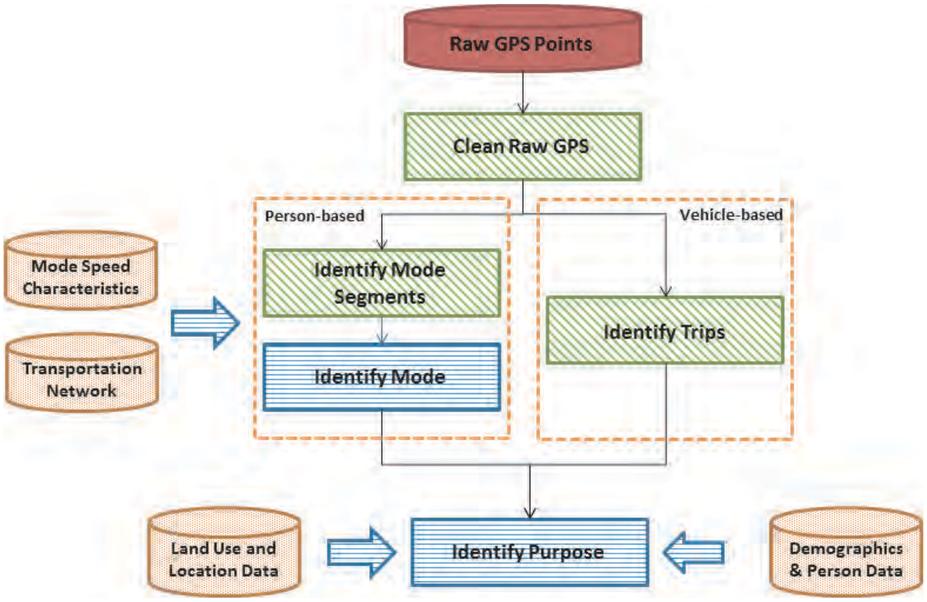


Figure 6. Procedures used to transform raw GPS point data into GPS trip data.

Inferring Travel Attributes

After the raw GPS traces have been cleaned and put into a standard format, travel behavior details can be extracted using additional steps (or algorithms). Figure 6 presents the overall process that transforms raw GPS point data into the travel attributes that are desirable for transportation planners and researchers interested in travel behavior. It also depicts the processing path used by person-based versus vehicle-based GPS trace data, as well as where the various extant data sources are used to help or improve the inference of specific travel attributes.

Individual trips are the building blocks of travel patterns. Trip end identification algorithms can be used to directly identify individual trips from the GPS stream with a high degree of spatial and temporal accuracy. Once trips are identified, they can be further processed and fused with ancillary data sets to derive location and type, travel mode, and trip purpose. Figure 7 provides an overview of the procedure used to infer travel attributes from single-trip GPS data.

Identifying Trip End Locations and Types

Over the past 15 years, several algorithms have been proposed to identify trips from a stream of GPS points. These algorithms review the trace data and identify locations where stops occurred by looking at points and speed values; the simpler approach is to discard all points not associated with movement and apply a minimum stop duration threshold (Wolf, Guensler,

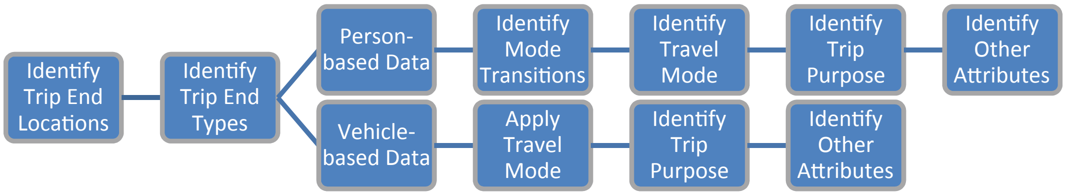


Figure 7. Procedure for inferring travel attributes from single GPS points.

and Bachman 2001). Another approach is to observe point density over time to identify activity clusters (Schüssler and Axhausen 2008).

Regardless of which approach is selected, it is important that the initial set of identified trip ends be reviewed for reasonableness. It is recommended that the same scrutiny that is typically applied to trip data obtained using household travel surveys be applied here. To make this possible, it is important to establish good data modeling and handling techniques, such as using standardized file formats or structured databases. Data visualization techniques can also be used to facilitate the process of reviewing the output of the trip end identification process.

After trip ends are confirmed, they can be classified using the usual travel demand location types: home, work, school, and other. If the GPS data in question were collected as part of an HTS, the usual way of performing this classification is to use the household, work, and school location data that were collected as part of the recruitment interview. The Trip Purpose section discusses how having data on household and personal locations is an important factor in the efficacy of trip purpose identification procedures.

If these location data are not available, then it is still possible to classify the identified stops using arrival and dwell time along with a few simple assumptions. For example, home is likely to be the place where the captured trips start and end in a given travel day, whereas work and school are likely to be places where the dwell time is the longest of all of the day's activities. The differentiation between work and school can then be done using person-level information such as age, school attendance, and employment.

Identifying Mode Transitions

When processing data collected using person-based devices, it is necessary to identify mode transitions (e.g., walk → bus, bus → walk), which may not be captured as part of the trip end identification process. Several methods have been proposed to perform this task, with the majority of them leveraging the speed and acceleration characteristics of these transition events and the proximity to public transport stops and routes. It is also important that the generated mode sequence be reasonable.

Tsui and Shalaby (2006) presented a method for detecting single-mode stages within trips by finding the points where the mode changed from walk to another mode or vice versa; the authors referred to these as mode transfer points (MTPs). Schüssler and Axhausen (2008) implemented a mode transition detection system based on the one proposed in Tsui and Shalaby (2006), with the original implementation featuring three types of MTP: end-of-walk (EOW), start-of-walk (SOW), and end-of-gap (EOG) points.

The EOG point was used to indicate the end of a period with GPS signal loss. For each transition from a speed below 2.78 m/s to above 2.78 m/s, the algorithm searches backward until the next point with a speed above 2.78 m/s or at least three consecutive GPS points with a maximum acceleration of 0.1 m/s^2 are found. In this case, the last of the trailing points with small acceleration values were marked as being potential EOW points; otherwise no EOW point was detected. The procedure for the potential SOW points follows the same logic, but in reverse order, while each point after a time difference of at least 80 s is marked as a potential EOG point.

As is true with any automated procedure, it is important to review the generated results for reasonableness based on known locations where mode transitions are likely to occur and probable mode sequences. Cases that fail these reasonableness tests should be reviewed by an analyst. Similarly to trip end reviewing, data visualization techniques can be used to simplify or improve this review process.

Identifying Travel Mode

Once trips are identified, travels modes can be identified. Most methods rely on speed-acceleration-deceleration profiles of various modes and the spatial relationships between the captured traces and the GIS transit and highway networks. Most methods will do an adequate job of classifying a trip as having been taken with a motorized versus nonmotorized travel mode. Travel modes that share the same infrastructure (i.e., urban bus and auto) are typically difficult to differentiate, and their effective identification may depend on the availability of GIS data on transit stops and segregated rights-of-way.

There are two main groups of methods that can be used to derive travel mode: rule-based and modeling or machine learning methods. Methods in the first group typically contain several thresholds, which may need to be calibrated to local conditions and may also rely on the presence of support data such as detailed GIS data on transit and rail facilities. However, they can be implemented without the need to first collect a model estimation or training data set. Based on the tests performed in Experiment A of Volume I, the fuzzy logic method (Tsui and Shalaby 2006, Schüssler and Axhausen 2008) will provide a good balance between error rates and data needs.

The second group depends on the presence of an existing data set that relates trip attributes to used travel modes. Their main advantage is that using data collected in the region where the method will be applied removes the need for calibration that exists in rule-based methods. Based on the tests performed in Experiment A of Volume I, the neural network machine learning method proved to be the best performer in this group of methods; however, it did struggle to differentiate between auto and bus modes. Thus, adding GIS data (i.e., proximity to bus stops at trip origin and destination locations) will likely improve the identification performance. Transit trips are also likely to be preceded by walk trips, although they may not always be captured by the GPS, and this information can also help further differentiate between auto and bus modes.

Finally, given the fact that speed and acceleration are the driving factors in travel mode identification, devices that can compute and report speed should be selected and minimum logging frequencies established. Whenever the technology selected or available permits (some devices may increase their power draw at higher logging frequencies, so tests should be done before a final selection is made), higher point logging frequencies (between 0.2 Hz and 1 Hz) should be selected, with the determination of the final logging frequency to be based on the devices' storage capacity and the desired logging rules. It is also worth noting that these higher frequencies will require longer download times.

Identifying Trip Purpose

The automatic identification of trip purpose (i.e., activity associated with the trip destination location) remains an issue that is difficult to solve. This is especially true if an attempt is made at classifying activities found in a GPS data stream into detailed trip purpose categories like those typically found in an HTS. If the full set of purpose categories is not needed for subsequent analysis and travel demand modeling, then a simplification of the purposes should be considered. Regardless, it is important to note that at the time this report was written, the methods available and tested were not sufficient for accurately estimating trip purpose for the range of purposes encountered in household travel surveys—including the methods contained in these guidelines.

There are two main groups of trip purpose identification methods. The first group relies on simple deterministic rules (Stopher, Clifford, and Zhang 2007) and is typically only capable of identifying major trip purposes, such as those associated with place types (i.e., home, work, school) and some of the other types based on the availability of frequently used locations for activities like shopping and banking. The second group of methods includes probabilistic and artificial intelligence models. These require more advanced analytical knowledge (statistical and mathematical), as well as the availability of a model estimation or calibration data set. If no data are available, then one can use models specified for similar use conditions, but this should be done while acknowledging that there may be challenges in data transferability.

Based on the tests performed for Volume I, it is clear that non-mandatory activities (i.e., shopping, eating out, maintenance, etc.) are hard to correctly identify. Higher-resolution spatial data could help disambiguate the competing choices in these purposes as well as provide more information on participants (e.g., collection of usual places used to perform discretionary and maintenance activities such as eating out, buying groceries, and banking). The tests also revealed that decision trees can be used to quickly generate working models and also to reveal relationships between the input data and selected purposes in the estimation data set. Information obtained in this manner can then be used to develop or refine choice models for trip purpose or to adapt existing heuristics-based methods.

Guidelines for Deriving Trip Purpose from Passive GPS Traces

The following set of suggestions was developed for practitioners who plan to use modeling to identify trip purpose using attributes that can be automatically derived from passively collected trace data and household and person characteristics:

- Research the availability of detailed land use and point of interest data for the study area and consider looking into commercial options.
- Ensure that the recruitment survey captures enough attributes to successfully classify all household members into a life-cycle category. This will reduce the number of assumptions made while preparing the data set for model estimation and deployment.
- Consider simplifying trip purpose classifications (e.g., combine shopping and maintenance purposes, or possibly all discretionary purposes)—this helps in the estimation of models and improves their success rate when applied to identify trip purposes. The goal here would be to define the minimum set of purposes that can be easily explained to survey participants in the calibration subsample and yet still provides enough resolution for travel demand modeling.
- Plan to collect personal locations (e.g., work, school, and volunteer) for each household member as part of the recruitment survey instrument. Build consistency checks into the GPS processing logic to ensure that captured mandatory locations are being matched to GPS destinations. A lack of a match in a typical travel day is unlikely and may be due to a poor geocode. Problems often arise when parking structures are situated away from the main building to which the geocode will typically fall closer to. This is especially important in mandatory purposes (work and school), given that the single most important driving variable for them was found to be a location match with the personal locations.
- Capture frequently visited locations (along with activities conducted at them) as part of the recruitment process. The availability of these data will assist in disambiguating choices between competing non-mandatory purposes—for example, grocery stores, ATMs, and gas stations.

Identifying Other Travel Attributes

Companions and shared travel can be easily identified if compatible trace data are available for most household members. In the recently concluded Cleveland GPS-based HTS, where participants 13 years old and older were instrumented with wearable loggers, companions on shared travel were identified based on a match of departure time from the previous place, arrival time, and end destination. Times were matched using a tolerance of 5 min, while coordinates were matched if they were within 75 m. The derived shared travel information was then put through the same logic check processes used in traditional diary-based HTSs. The process worked fairly well, with most problems occurring when multiple short-duration activities had occurred (i.e., within the 5-min tolerance used) and caused mismatches between the data of different participants.

Inferring Tours and Activity-Travel Patterns

The construction of individual travel tours from GPS traces follows the process of inferring trips and trip attributes. The ability to infer tours requires a unique link between individual trips, as discussed previously. The inference of full daily activity-travel patterns takes the inferred trips and trip characteristics as input and aggregates these individual trips into tours and potentially into activity patterns if activity data were inferred in the previous step. The identification of tours is very straightforward in the absence of any activity or location information being inferred in the previous step. In this case, it is only necessary to identify when the trips return to a previously visited location, which closes the tour. However, this type of analysis does not provide much in the way of useful information, so the remainder of this section will discuss identifying tours in the context of a daily travel pattern. Figure 8 provides an overview of the process used to infer tours and activity-travel patterns.

Data Requirements for Tour Identification

Identifying tours within a daily activity-travel pattern requires additional information regarding the input data and the estimated data resulting from the trip inference step. The additional requirements include the identification of the home, work, and school locations for each individual, and the time spent at activities that occur at the trip ends. If the identified locations in the GPS traces can be further classified with an activity type, this leads to more detailed activity-travel pattern information but is not strictly required. If activity types cannot be identified, it is sufficient to leave the locations that are not home, work, or school as a general "other" category. However, the home, work, and school locations are necessary to identify tours and sub-tours in a daily travel pattern context. The start time and duration of each trip should be carried forward here from the trip identification routine. Also, if the data collection encompasses more than 1 day, the date on which the travel occurred should be used to generate the multiday travel patterns.

The required home, work, and school locations should be determined in the trip inference stage. However, if this is not done, they can be inferred from the trip data, provided the data represent a full day, and ideally represent several days. The home location can generally be assumed to occur where the trips start and stop for the day, while the work location (if it exists) can be identified

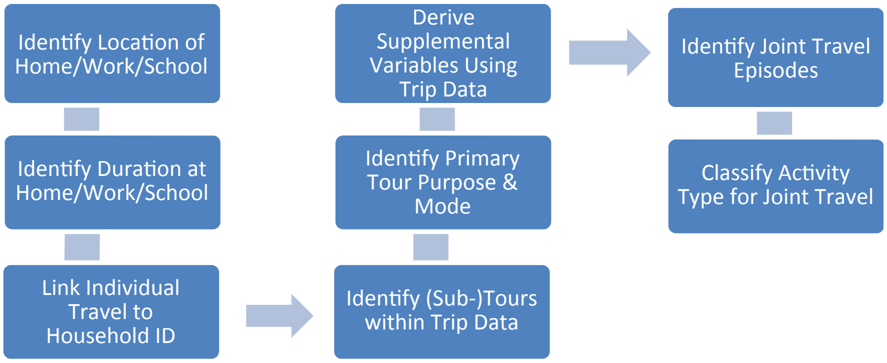

Identify Location of Home/Work/School
Identify Duration at Home/Work/School
Link Individual Travel to Household ID
Identify (Sub-)Tours within Trip Data
Identify Primary Tour Purpose & Mode
Derive Supplemental Variables Using Trip Data
Identify Joint Travel Episodes
Classify Activity Type for Joint Travel

Figure 8. Procedure for inferring tours and activity-travel patterns from GPS data.

Table 4. Minimum data requirements for tour identification.

Variable	Data Type	Description
Basic Requirements		
Person number	Unique identifier	Required identifier to link trips
Location ID	Unique identifier	Unique identifier for physical location
MODE	Integer 1–10	Walk, bike, drive, pass, transit, paratransit, taxi, school bus, carpool
TRPDUR	Integer	Trip duration
Daily Travel Pattern Analysis		
Location type	String	Required location types: home, work, school, other
ACTDUR	Integer	Activity duration – time spent at trip end
Start time/day	Integer	Start time of the activity
Activity type	Integer	Optional: type of activity
Household-Level Analysis		
HH number	Unique identifier	Required if doing household-level analysis, otherwise optional

somewhat reliably by observing long-duration activities that are away from the home location. The reliability here improves with the duration of data collected since the workplace tends to be a routinely visited location for employed individuals. The school location can be identified in a similar manner, and the reliability here can be improved by utilizing databases of school locations, such as those that can be obtained from state boards of education or the National Center for Education Statistics.

Finally, if the individual travelers can be linked to households (i.e., in a controlled sample such as a household GPS survey were all members are tracked), the household ID can be included. The addition of household-level information makes it much easier to identify certain activity types such as picking up/dropping off and serving dependents. It also allows for joint activity-travel information to be estimated. Outside of providing more detailed activity-travel information, the use of household-linked GPS traces also greatly enhances the ability to derive socio-demographic information from the data. Unfortunately, the household information is not often available from mass data sources, so in these cases the analysis is limited to the person level. The data requirements discussed here are summarized in Table 4.

Tour Identification Process

When the required input data in the form of trips have been obtained, a simple tour identification process can be used to produce the daily activity-travel patterns, from which a variety of travel pattern indicator variables can be calculated for later use in the demographic inference procedure.

The tour identification starts by parsing all trips for an individual and identifying the anchor points, which are generally the home, work, and school locations. These points serve as the starting and ending points for tours, with main tours starting/ending at home and sub-tours starting/ending at work or school by convention. The trips are then formed into consistent tours and sub-tours by observing the anchor points. To accomplish this, the trips for each individual are first sorted by the start time of the trip. Then, each trip is checked to determine if it is a home location anchor point, and if so, a new tour is started and trips are added until another trip is found destined for the home location.

It should be noted that it may be necessary to convert trip-based GPS data into a place sequence, which in most cases can be done by adding a home location observation to the start of the data. During this process, if a work/school anchor point is found, a potential sub-tour is started, and subsequent acts are added to the potential sub-tour. If the same work/school anchor is found again, then the sub-tour is closed and added to the main tour. However, if no additional work/school anchor is found before reaching home again, the potential sub-tour is removed and the trips are added to the main tour, which is then closed. This process repeats until all trips have been visited. If multiday traces have been collected, then the artificial start-of-day anchor points do not need to be added for subsequent days as long as the previous day terminated at home. Figure 9 presents a visual overview of the tour identification process.

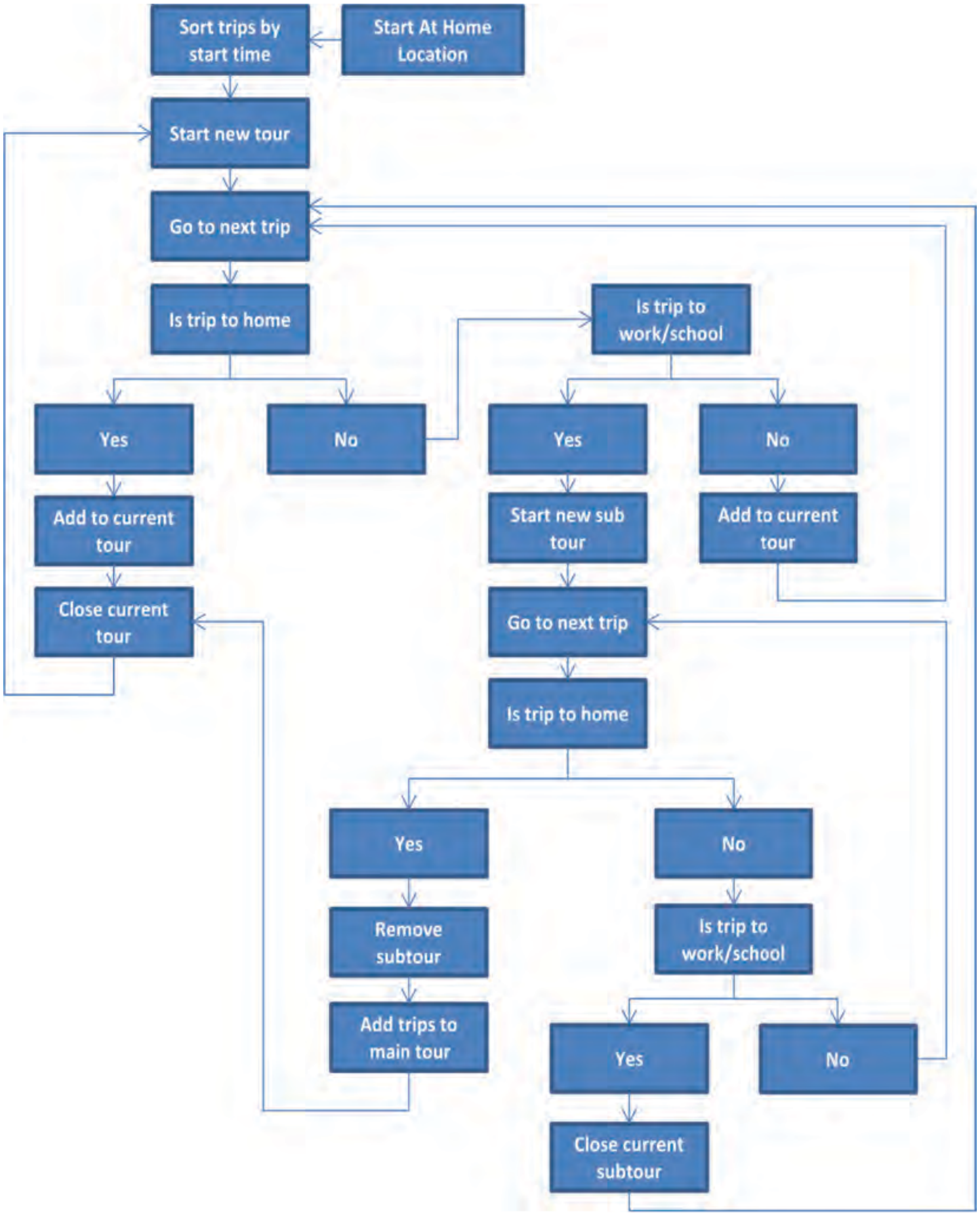


Figure 9. Tour identification process.

Table 5. Variables that can be estimated/derived from tour data.

number of total tours per day	number of work activities
number of sub-tours per day	total duration of all work activities
number of work tours	average duration of work activities
number of school tours	number of school activities
average number of stops per tour	total duration of all school activities
average number of stops per work tour	average duration of school activities
average number of stops per school tour	number of pickup/drop-off activities
average number of stops per other tour	total duration of all pickup/drop-off activities
average travel time per tour	average duration of pickup/drop-off activities
average travel time per work tour	number of other activities
average travel time per school tour	total duration of all other activities
average travel time per other tour	average duration of other activities
total time spent at home	percentage of tours by auto mode
percentage of work tours by auto mode	percentage of school tours by auto mode

Guidelines for Deriving Activity-Travel Pattern Characteristics from GPS Traces

Once each tour has been identified in the trip data, a large number of activity-travel pattern characteristics can be derived using this information. These variables are calculated by parsing the trips that have been associated with each tour. The most important variables to identify are the primary tour purpose and the primary tour mode, which are important for travel demand analysis and also for the demographic inference process. These characteristics are identified by observing the trips associated with each tour. Tours with a work or school activity generally have that activity considered as the primary purpose. For tours with other activity types (if available), the primary purpose can be identified by observing the activity with the greatest duration, by applying activity priority rules, and so forth. Similarly, the primary tour mode can be identified as the mode with the longest travel time, or precedence rules can be used such as transit supersedes auto, which supersedes walk, and so forth. Once the basic characteristics have been identified, a wealth of supplemental variables can be calculated, even for the most basic input case described here. Some examples of variables that can be derived are shown in Table 5.

The tour variables listed can be extended if household linking is included in the GPS traces. Household linking allows the identification of joint travel episodes, whether fully joint or partially joint, and supplementation of the variables described previously with joint travel characteristics. The use of joint information allows the identification of pickup and drop-off activities, which are significant for identifying demographics in the next stage. However, in the general case, such information is not available.

Inferring Demographics

Inferring some amount of individual or household-level demographic information based on the identified tour patterns and trip information can be done after the inference of the tours and travel patterns. This can be done in a number of ways; however, the methodology identified as the most promising involves applying models developed using existing household travel surveys to the observations derived from GPS trace data. These models represent, in effect, an inverse travel demand model, whereby activity-travel pattern characteristics are used to estimate demographics, contrary to the traditional travel demand modeling practice. A straightforward framework for developing these models, as well as an example set of models that can be used to infer demographics estimated from Chicago survey data, will be presented briefly in this section. Note that the demographic models presented in these guidelines can be re-estimated using the same process with different source data. Also note that the demographic characterization of samples based on travel characteristics is a more experimental methodology than that discussed in Experiment A, and it should be used with caution. It is not intended that the methodologies discussed in this section be used to replace demographic questionnaires in traditional household surveys in the same way that the methods discussed in Experiment A can replace travel characteristic questions in surveys. These methods are, at this point, intended solely to demonstrate how demographic characteristics can be inferred for anonymous sources of travel data in which other sources of demographic data do not exist.

Demographic Estimation Model Development Process

The recommended process flow for developing demographic inference models is shown in Figure 10. The procedure follows the derivation of the person-level travel and tour characteristics discussed previously. The demographic characterization procedure for the GPS traces involves two stages. First, aggregate-level person-type clusters are developed for later demographic modeling. These clusters serve as the dependent variable for the later stages, so that in the personalization process, the major type of the person is selected first, then, depending on the type of the person, more detailed demographics are modeled. In the second stage, the travel pattern data and local land use data are used to select one of the major person types. The primary person type is then used, along with the travel characteristics and any other land use information, in a set of models to determine other socio-demographic characteristics. It is important to note here that this model procedure is intended as a current guideline for the estimation of demographic characteristics from GPS traces and that any of the models could and should be re-estimated following these general guidelines.

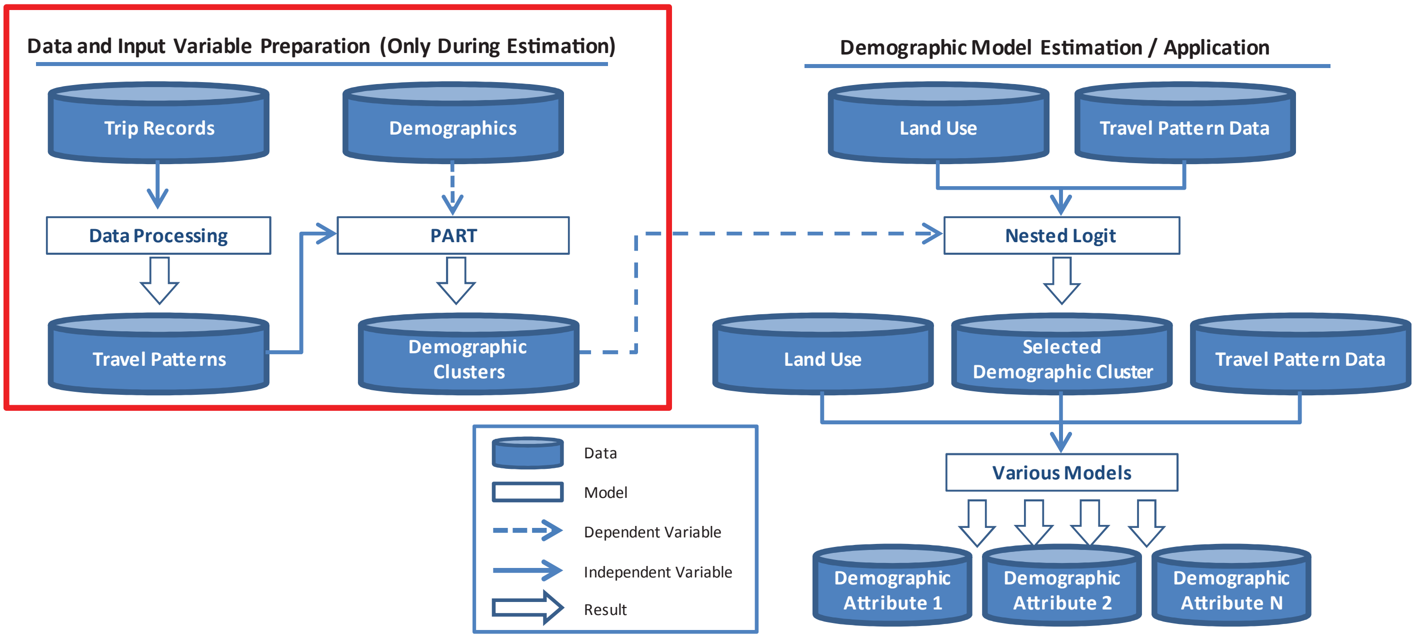


Figure 10. Demographic model estimation and application procedure.

Table 6. Primary person types used in demographic modeling.

Class	Description
1	Part-time workers
2	Full-time workers
3	Retirees
4	Young children
5	School children
6	Others

The first task in the procedure is the development of the primary demographic clusters as shown in Figure 10. In the completion of the demographic cluster estimation, several subtasks are necessary. First, the tour and daily activity pattern variables extracted from the survey data are transformed into a set of primary factors using a principal components analysis, given the high levels of interdependence that existed between many of the variables. These factors are then clustered using the *K*-means clustering algorithm. Finally, the major person demographic types are developed using a partial decision tree classification algorithm (PART), with the travel pattern cluster membership distribution serving as the univariate dependent variable.

The process selects the major person types with the maximum differences in travel cluster membership distribution (i.e., the person types with the most differences in travel patterns). This stage is shown in the box in the upper left of Figure 10. The clusters developed here represent the minimal set of person demographics that best determine activity-travel pattern differences at a high level. These clusters are very unlikely to change when different input data are used since they closely align to traditionally used person types in travel demand analysis. Therefore, it is suggested that these person-type clusters be used when estimating new demographic models, effectively removing this step from the model estimation process.

A description of the most useful demographic clustering model found in the Volume I analysis is shown in Table 6. These clusters serve as the primary dependent variable for the demographic models.

A variety of models have been used to classify individuals in terms of five primary person-level attributes: primary person type, educational attainment, gender, possession of a driver's license, and age. The individuals observed in the GPS traces are also classified as belonging to a set of household types, which incorporate the household size, number of vehicles, and presence of children dimensions. The final models were selected for performance reasons but also to give a representation of both joint and non-joint modeling methods and regression-type models versus machine learning-based approaches. It should be noted that there is generally very little difference in results obtained from machine learning compared to the discrete choice modeling approaches. The machine learning models are usually easier to estimate and apply, while the discrete choice models generally handle the modeling of joint characteristics and correlated characteristics better. Therefore, it is suggested that the C4.5 decision tree be used when estimating single characteristics and in situations where the sensitivity of the outcome to various variables is not of interest. Discrete choice models should be used when the elasticities are of interest for policy reasons or when estimating characteristics jointly. An overview of the modeling procedures is shown in Table 7.

Table 7. Models used for demographic characterization.

Variable	Values	Model Procedure
Person/worker status	Part-time worker, full-time worker, retiree, child, schoolchild, other person type	Nested logit
Educational attainment	No high school, high school, college	*with above*
Age	0–16, 16–25, 25–45, 45–65, 65+	Ordered Logit
Gender	Male, female	Binary logit
Licenses	Yes, no	PART decision list
HH size	1, 2, 3+	C4.5 decision tree
Number of vehicles	0, 1, 2+	*with above*
Has children	No, yes	*with above*

Demographic Characterization Model Usage

The demographic characterization process outlined here should be used with caution. However, this process appears promising given that the models generally showed substantial improvement over null model expectations, appeared to be transferable to some degree, and were able to generate consistent person-characteristic estimates. These results are significant in light of the minimal data used as input to the personalization process. Thus, as more data from detailed land use databases become available and larger and longer-duration GPS trace collection procedures are developed, the models could be improved substantially. The models listed previously have been estimated using the Chicago Household Travel Survey data and are available in Appendix E of Volume I.

The models described in these guidelines can be used directly to assign demographic characteristics to GPS traces; however, it is suggested that the models be re-estimated following the steps described previously, using local data if available, or using national data (i.e., the national Household Travel Survey) before applying to the anonymous GPS traces. The model can also be updated with locally collected samples if the data available for full model re-estimation are insufficient. The local updating procedure can be handled in several ways. The most straightforward method is to apply the model to a small, local sample where both the travel patterns and demographics are known and then adjust the model constants until the estimated demographics and observed demographics match. The calibrated model could then be applied to local GPS data.

Limitations of the Demographic Characterization Process

If this process is to be used, it is important to be aware of several limitations and pitfalls associated with the demographic characterization process. Several key findings, which were observed in the development of the model, can help improve the application of such a procedure and can help to guide the data collection process used to gather the input GPS traces.

- Multiday data collection is very preferable to single day data collection since it helps average out intrapersonal day-to-day variation, which can be greater than interpersonal variation. This variability tends to confound the personalization procedure (e.g., if a worker is surveyed on a nonworking day, the person will be difficult to identify as a worker). This finding aligns with previous research (Pas and Sundar 1995).
- It is also important to have, at a minimum, reasonable estimates of workplace and school locations, either from access to more detailed location databases or from longer-term observation that can identify working patterns.
- If it were possible to ensure that all household members were tracked and linked together, then more accurate estimates of certain person and household characteristics could be made. This would especially help since the joint trip-making travel characteristics tended to be significant in previous versions of the model effort, which did not conform to Experiment A.
- The causality between travel patterns and personal characteristics here runs counter to how the modeling is usually done, and as such appears to be much weaker than in the reverse case [i.e., people travel the way they do because of who they are, but people are generally not who they are because of how they travel (with some exceptions, such as with possession of a license or vehicle ownership)].
- A significant problem is that some person types are virtually indistinguishable based on travel characteristics alone (e.g., the young child travel pattern, naturally, looks much like the caretaker's pattern). This is especially true for short-term data collection. For example, part-time and full-time workers are rarely distinguishable in the short term (most part-time workers work full shifts for fewer days).

- Finally, the joint modeling of attributes is difficult but provides some benefit to modeling separately, outside of just consistency, which is itself important.

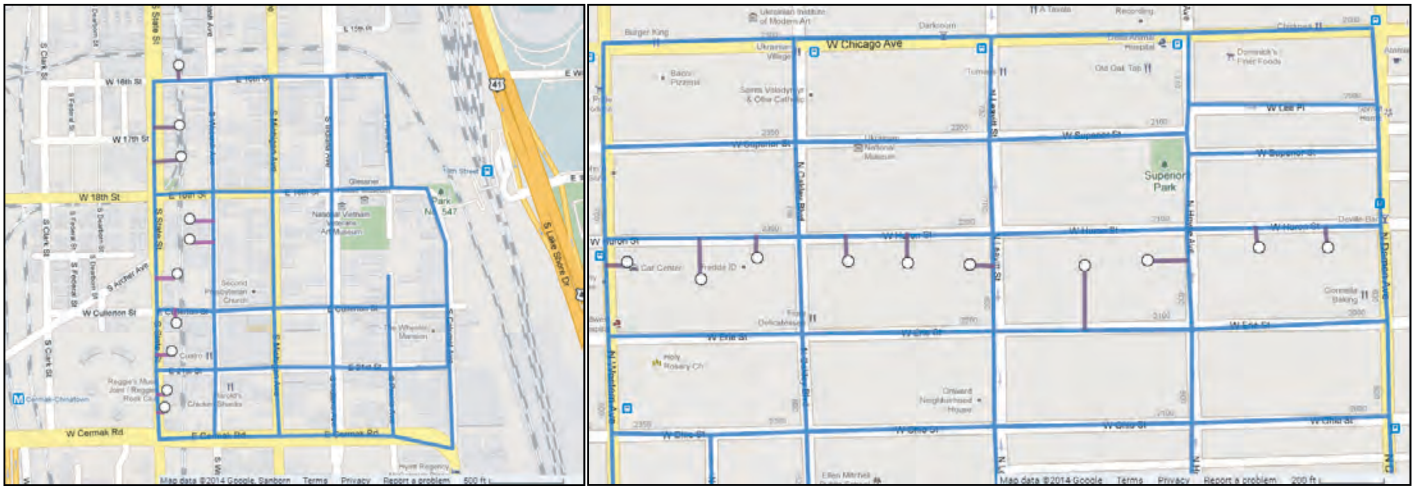
Following the procedures outlined here, and keeping in mind these findings when selecting data sources and developing or applying the models, should allow for a fairly robust estimation of demographic characteristics.

Map Matching

Independent of the previous procedures listed is a frequently desired need to map match GPS data to a digital road network. Successful map matching allows planners and researchers to translate trip characteristics revealed in the GPS trace to uniform road databases frequently used for modeling or other planning purposes. Further, map matching the GPS data builds a connection between trips and other GIS attributes that exist or can be related to roads. Inferring travel routes across a road network from GPS data is a three-step process requiring the previously discussed data cleanup, the acquisition of a digital representation of the road network, and map matching between the GPS data and the road network. The final step is only possible after the first two are complete.

For map matching to work at all, a digital map of the study area must be available so that GPS traces can be matched to specific points on the local transportation network. Road databases suitable for map matching can be acquired from private companies, but many are freely available from open sources or public agencies. While an accurate and timely digital road network will make processing considerably simpler, there are also approaches to handle inaccuracies in maps, which may arise even in professional maps due to error, the use of nonpublic roads, or the passage of time.

The map matching of GPS traces to network links is a well-documented research subject. It is not as simple a process as identifying the nearest line for each point. Figure 11(a) shows an exaggeratedly inaccurate trace indicated with white circles. In this image, the short lines attached to these points are attached to the matched link based on nearest road segment. The resultant ordered set of links is jumbled, listing four different streets for the straight line path. It is also an example of an incomplete digital network. While the inaccurate trace could represent the westernmost north–south street, it aligns better with the passenger rail line half a block to the



(a) Point-to-Curve Algorithm

(b) Topological Algorithm

Figure 11. Examples of weak map-matching algorithms.

east. Figure 11(b) displays the work of a topological algorithm that takes the junctions of streets into consideration when selecting routes, choosing only from those that share a node with the current line segment. While it results in a logically ordered path, it does not result in the correct one. Again, the simplistic assumption that the nearest line segment is the correct one results in problems near intersections. For this reason, advanced map-matching algorithms account for uncertainty at each point by taking into consideration a series of points rather than individual points. This could range from a few points to all of the points in the trace.

Independently developed, advanced map-matching algorithms typically must consider past and future points for each point they assign to a link. Points may not be permanently assigned to links until a few steps after initial deductions. One such advanced algorithm, which was developed at the University of London, searches for the nearest network node to consider the set of links extending outward from it. Conditional tests are then conducted to finalize the initial choice and locate its progress along the segment (Quddus, Noland, and Ochieng 2006). Another successful example continuously evaluates the *N*-most likely routes based on the set of points. After each point is analyzed, each of the *N* carried routes is evaluated to develop a score and ranking. As new links are suggested, they are combined with the existing set of likely links to develop a new set of *N*-most likely routes (Marchal, Hacknet, and Axhausen 2005; Schüssler and Axhausen 2008; Pereira, Costa, and Pereira 2009).

The route link data may be used to summate the link-level characteristics of individual routes, resulting in mileage or frequency data for such occurrences as highway driving, toll use, express route use, and neighborhood preference or disfavor. The total of these route traits may be used to train a model of route choice decisions that can estimate a preferred route given an origin, destination, and user. Typical inputs are travel time, travel distance, number of left turns, cost, and highway distance. The resulting travel routes should be more precise and greater in number than those collected by travel diaries.

Privacy Considerations

Overview of Private Information

GPS location data can clearly identify where the data source is throughout the time period that GPS data are collected. Whether collected by individuals or vehicles, intermittent location information or second-by-second GPS trace data can pinpoint home, work, school, and other frequently visited locations. These locations serve to uniquely identify the person from whom the data were collected. Thus, these types of GPS data are considered to be personally identifiable information. Furthermore, data that are collected from consumer devices and resold to transportation planners often originate from, or possibly contain, personally identifiable information. Legitimate efforts to protect private information by data collectors, integrators, and users have met with limited success since investigative efforts can sometimes reveal the underlying identity behind the GPS trace (Elango et al. 2013).

In travel behavior research and transportation planning, detailed travel data regarding individuals are valuable for understanding travel choices and the building of complex transportation models. Providing full protection of privacy while maintaining the full usefulness of the data is difficult, if not impossible, to achieve (Ohm 2010). Private information, as considered here, is any data element that can be used to identify a specific person or household. This would include direct attributes (names, addresses, etc.), detailed time and positional information regarding activities (full GPS traces, geocodes), and information that can be connected to other data sources (employment records, consumer data). Confidential information includes stated or revealed travel choices from an individual or family. The following sections discuss privacy/confidentiality considerations with GPS data collected within surveys of personal mobility or from consumers at large.

Researchers and planners that use private data in some manner are encouraged to review and adopt the policies described in Title 45, Code of Federal Regulations Part 46, Protection of Human [Research] Subjects. This legislation, designed originally for health research, provides specific definitions and data management guidelines for the use and storage of private information.

GPS Data Collected in Surveys

Be their nature, surveys involve multiple levels of consent—for participation and data provision. Households contacted to participate in the two-stage HTS can elect to participate or withdraw at each stage of the survey—before the recruitment survey, after the recruitment survey, or during the travel reporting/retrieval survey.

Due to declining landline telephone usage by households (lines that were previously used to recruit via random digit dialing), travel survey recruitment today is typically initiated using a listing of U.S. Postal Service delivery addresses for a given survey area. Since the names of the persons living at these addresses are not available in these listings (and therefore are not known), initial materials inviting survey participation are addressed to "Resident." Households located at targeted addresses are invited to participate in the HTS and can elect not to respond to the survey request. Until the time at which the household completes the initial recruitment survey, the names of household members are not known by the survey firm nor are they associated with the sampled address. This is intentional since address-based sampling is intended to sample based on addresses, not names associated with addresses (and often the names that may be associated and available for a given address are incorrect or dated).

Given the mixed survey mode (telephone and web) data collection methods that are now typical in household travel surveys as well as in other surveys, it is unusual for written consent to be requested for participation. Instead, participants are asked during the recruitment process if they are willing to participate in the survey, and survey materials often include a toll-free number that participants can call with any questions. Recruited households are told that participation is voluntary, that they can choose to opt out of the survey at any time, and that their personal information will be kept confidential, as required by law.

GPS data collected in travel surveys are provided to clients and other stakeholders without including any directly identifiable information. Regardless, the raw GPS data can be processed (as discussed in this document) to identify the home, school, and work locations of the households that collected the data. These GPS data are typically uploaded to secure, password-protected FTP sites that are only accessible to data users who have a signed confidentiality agreement with the client. The data are posted to these sites for a limited amount of time to ensure the utmost security.

In rare cases, the agencies sponsoring GPS-enhanced or GPS-based surveys may mask or modify the latitude and longitude coordinates of participant home, work, and school locations to alter the exact location. When this is done, a geographic unit [such as a fixed-grid overlay or a predefined unit such as a census block or traffic analysis zone (TAZ)] is often chosen to be the assigned location of all trip ends originating or ending in that unit. Other agencies have chosen not to provide any location information (addresses or coordinates) in public-access data sets to ensure privacy protection for participants.

As mentioned previously, some state and regional agencies are electing to provide GPS travel survey data to the TSDC, which is operated by the National Renewable Energy Laboratory (NREL), for researchers to access and analyze. This secure data center was created to meet the FHWA's desire to have a permanent central repository of GPS data sets collected in a range of transportation studies or surveys and to provide access to these data sets to researchers and practitioners in a highly secure manner that maintains the privacy of individual surveyed households. More information about this center can be found at http://www.nrel.gov/tsdc.

GPS Data Collected (or Crowd-Sourced) from Consumers

Private data vendors rely on raw data sources that come from fixed sensors, commercial fleet vehicles, and private consumers. Data from commercial fleets are subject to data licensing agreements that focus on reselling the data and are not generally concerned with the privacy protection of the commercial vehicle drivers. GPS data from private consumers are gathered through in-vehicle or mobile consumer devices. These hardware and software products come with licensing agreements that indicate that certain information may be shared with others. For example, Google offers the following statements as part of its end-user license agreement:

> When you use a location-enabled Google service, we may collect and process information about your actual location, like GPS signals sent by a mobile device. We may also use various technologies to determine location, such as sensor data from your device that may, for example, provide information on nearby Wi-Fi access points and cell towers.
>
> We may share aggregated, non-personally identifiable information publicly and with our partners—like publishers, advertisers or connected sites. For example, we may share information publicly to show trends about the general use of our services.

The broad licensing agreements that are accepted by users allow devices and applications freedom to use GPS trace data, other location data, and social network information, with the stated intent of providing the user with better services. This data mining approach can provide companies with very detailed profiles of individuals that can be used for marketing purposes (Gralla, Sacco, and Faas 2011). Additionally, and of concern for this report, the packaging and reselling of this data to researchers and planners is possible. While license agreements may protect a company using private information, individuals and organizations funded through public sources should consider carefully their responsibilities when using such data.

The concern over privacy is an ongoing issue with consumer-generated data products, and this topic could be heavily influenced by future legislation. In some cases, there may be specific legal restrictions on the use of private consumer data by public officials. In 2012, the State of California passed the Location Privacy Act that required law enforcement organizations to secure a search warrant before accessing personal travel data. There are many who believe that the same sort of restriction should be placed on other public uses of private data. The Geolocation Privacy and Surveillance Act is a bipartisan bill that has been under consideration by the U.S. Congress since 2011. If passed and implemented, it would dictate clear guidelines for the use of GPS data collected from private consumers. The act as written is also clear that specific consent is needed and that public agency restrictions are limited to only investigative or law enforcement purposes.

As described in previous sections, there are several techniques used by private companies to protect privacy. These data aggregations and perturbations generally accomplish the desired level of protection. As the market value for more detailed information (individual trip traces) increases, private companies may respond with more relaxed protections.

References

Asakura, Y., and E. Hato. 2004. "Tracking Survey for Individual Travel Behaviour Using Mobile Communication Instruments." *Transportation Research Part C*, Vol. 12, No. 3–4, pp. 273–291.

Bachman, W., M. Oliveira, J. Xu, and E. Sabina. 2012. "Using Household-Level GPS Travel Data to Measure Regional Traffic Congestion." Presented at the 91st Annual Meeting of the Transportation Research Board. Washington, D.C.: Transportation Research Board of the National Academies.

Battelle Memorial Institute. 1997. Lexington Area Travel Data Collection. Washington, D.C.: Federal Highway Administration.

Cambridge Systematics, Vanasse Hangen Brustlin, Inc., Gallop Corporation, C. R. Bhat, Shapiro Transportation Consulting, LLC, and Martin/Alexiou/Bryson, PLLC. 2012. *NCHRP Report 716: Travel Demand Forecasting: Parameters and Techniques.* Washington, D.C.: Transportation Research Board of the National Academies.

Christidis, P. and N. Ibanez Rivas. 2013. "Road Congestion in Europe: Measurement and Monitoring Using GPS Traces." Presented at the 92nd Annual Meeting of the Transportation Research Board. Washington, D.C.: Transportation Research Board of the National Academies.

Elango, V. and R. Guensler. 2011. "On Road Vehicle Activity GPS Data and Privacy." Presented at the 13th National Transportation Planning Applications Conference. Portland, OR: Transportation Research Board of the National Academies.

Elango, V., S. Khoeini, Y. Xu, and R. Guensler. 2013. "Longitudinal GPS Travel Data and Breach of Privacy via Enhanced Spatial and Demographic Analysis." *Transportation Research Record: Journal of the Transportation Research Board No. 2354.* Transportation Research Board of the National Academies, pp. 86–98.

Gralla, P., A. Sacco, and R. Faas. 2011. "Smartphone Apps: Is Your Privacy Protected?" *Computerworld*, July 7, 2011.

Gur, J., B. Shlomo, C. Solomon, and L. Kheifits. 2009. "Intercity Person Trip Table for Nationwide Transportation Planning in Israel Obtained from Massive Cell Phone Data." *Transportation Research Record: Journal of the Transportation Research Board, No. 2121.* Washington, D.C.: Transportation Research Board of the National Academies, pp. 145–151.

Lawson, C., C. Chen, and H. Gong. 2010. Advanced Applications of Person-Based GPS in an Urban Environment. Albany, NY: University at Albany – State University of New York.

Liu, H., Y. Wang, and X. Chen. 2013. "Vehicle Emission and Near-Road Air Quality Monitoring in Shanghai, China, Based on Taxi GPS Data and MOVES Revised Emission Inventory." *Transportation Research Record: Journal of the Transportation Research Board, No. 2340.* Washington, D.C.: Transportation Research Board of the National Academies

Ma, J., F. Yuan, C. Joshi, H. Li, and T. Bauer. 2012. "A New Framework for Development of Time Varying OD Matrices Based on Cellular Phone Data." Presented at the 4th Annual Conference on Innovations in Travel Modeling. Tampa, FL: Transportation Research Board of the National Academies.

Marchal, F., J. Hacknet, and K. Axhausen. 2005. "Efficient map-matching of large GPS data sets–Tests on a speed monitoring experiment in Zurich." Zurich, Switzerland

National Research Center Inc. 2007. *Trip Diary Survey of Community Travel Patterns. Report of Results.* Flagstaff, AZ: Flagstaff Metropolitan Planning Organization.

Ohm, P. 2010. "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization." *UCLA Law Review*, Vol. 57, Issue 5, pp. 1701–1777.

Pas, E. and S. Sundar. 1995. "Intra-Personal Variability in Daily Urban Travel Behavior: Some Additional Evidence." *Transportation*, Vol. 22, pp. 135–150.

Pereira, F., H. Costa, and N. Pereira. 2009. "An Off-Line Map-Matching Algorithm for Incomplete Map Databases." *European Transport Research Review 1*, pp. 107–124

Quddus, M. A., R. B. Noland, and W. Y. Ochieng. 2006. "A High Accuracy Fuzzy Logic Based Map Matching Algorithm for Road Transport." *Journal of Intelligent Transportation Systems*, Vol. 10, No. 3, pp. 103–115.

Ratti, C., R. M. Pulselli, S. Williams, and D. Frenchman. 2006. "Mobile Landscapes: Using Location Data from Cell Phones for Urban Analysis." *Environment and Planning B*, Vol. 33, No. 5, pp. 727–748.

Schüssler, N., and K. W. Axhausen. 2008. "Identifying Trips and Activities and Their Characteristics from GPS Raw Data Without Further Information." Presented at the 88th Annual Meeting of the Transportation Research Board. Washington, D.C.: Transportation Research Board of the National Academies.

Stopher, P., E. Clifford, and J. Zhang. 2007. "Deducing Mode and Purpose from GPS Data." Presented at the 11th National Transportation Applications Planning Conference, Daytona Beach, FL: Transportation Research Board of the National Academies.

Stopher, P., Q. Jiang, and C. Fitzgerald. 2005. "Processing GPS Data from Travel Surveys." In Proceedings from the 2nd International Colloquium on the Behavioural Foundations of Integrated Land-Use and Transportation Models: Frameworks, Models and Applications, PROCESSUS, Toronto, Canada.

TTI and Lee Engineering. 2011. Private Probe Vehicle Data for Real-Time Applications-Final Report. Maricops Association of Governments.

Tsui, S. Y. A., and A. Shalaby. 2006. "An Enhanced System for Link and Mode Identification for GPS-based Personal Travel Surveys." *Transportation Research Record: Journal of the Transportation Research Board, No. 1972.* Washington, D.C.: Transportation Research Board of the National Academies. pp. 38–45.

Wolf, J. 2004. "Defining GPS and Capabilities." In *Handbook of Transport Geography and Spatial Systems* (K. J. Button, K. E. Haynes, P. R. Stopher, and D. A. Hensher, eds.). Oxford, UK: Elsevier.

Wolf, J., R. Guensler, and W. Bachman. 2001. "Elimination of the Travel Diary: An Experiment to Derive Trip Purpose from GPS Travel Data." Presented at the 80th Annual Meeting of the Transportation Research Board, Washington, D.C.:Transportation Research Board of the National Academies.

Wolf, J., M. Oliveira, and M. Thompson. 2003. The Impact of Trip Underreporting on VMT and Travel Time Estimates: Preliminary Findings from the California Statewide Household Travel Survey GPS Study. *Transportation Research Record: Journal of the Transportation Research Board, No. 1854.* Washington, D.C.: Transportation Research Board of the National Academies. pp. 189–198.

Xu, Y., V. Elango, R. Guensler, and S. Khoeni. 2013. "Idle Monitoring, Real-Time Intervention, and Emission Reductions from Cobb County School Buses." *Transportation Research Record: Journal of the Transportation Research Board, No. 2340.* Washington, D.C.: Transportation Research Board of the National Academies.

Zito, R., G. D'Este, and M. A. P. Taylor. 1995. "Global Positioning Systems in the Time-Domain–How Useful a Tool For Intelligent Vehicle-Highway Systems." *Transportation Research Part C–Emerging Technologies* 3, No. 4, pp. 193–209.

Abbreviations and acronyms used without definitions in TRB publications:

A4A	Airlines for America
AAAE	American Association of Airport Executives
AASHO	American Association of State Highway Officials
AASHTO	American Association of State Highway and Transportation Officials
ACI–NA	Airports Council International–North America
ACRP	Airport Cooperative Research Program
ADA	Americans with Disabilities Act
APTA	American Public Transportation Association
ASCE	American Society of Civil Engineers
ASME	American Society of Mechanical Engineers
ASTM	American Society for Testing and Materials
ATA	American Trucking Associations
CTAA	Community Transportation Association of America
CTBSSP	Commercial Truck and Bus Safety Synthesis Program
DHS	Department of Homeland Security
DOE	Department of Energy
EPA	Environmental Protection Agency
FAA	Federal Aviation Administration
FHWA	Federal Highway Administration
FMCSA	Federal Motor Carrier Safety Administration
FRA	Federal Railroad Administration
FTA	Federal Transit Administration
HMCRP	Hazardous Materials Cooperative Research Program
IEEE	Institute of Electrical and Electronics Engineers
ISTEA	Intermodal Surface Transportation Efficiency Act of 1991
ITE	Institute of Transportation Engineers
MAP-21	Moving Ahead for Progress in the 21st Century Act (2012)
NASA	National Aeronautics and Space Administration
NASAO	National Association of State Aviation Officials
NCFRP	National Cooperative Freight Research Program
NCHRP	National Cooperative Highway Research Program
NHTSA	National Highway Traffic Safety Administration
NTSB	National Transportation Safety Board
PHMSA	Pipeline and Hazardous Materials Safety Administration
RITA	Research and Innovative Technology Administration
SAE	Society of Automotive Engineers
SAFETEA-LU	Safe, Accountable, Flexible, Efficient Transportation Equity Act: A Legacy for Users (2005)
TCRP	Transit Cooperative Research Program
TEA-21	Transportation Equity Act for the 21st Century (1998)
TRB	Transportation Research Board
TSA	Transportation Security Administration
U.S.DOT	United States Department of Transportation